UNCERTAINTY ESTIMATION FOR TARGET DETECTION SYSTEM DISCRIMINATION AND CONFIDENCE PERFORMANCE METRICS

DISSERTATION

David R. Parker, Major, USAF

AFIT/DS/ENG/06-01

DEPARTMENT OF THE AIR FORCE

AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright Patterson Air Force Base, Ohio

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/DS/ENG/06-01

UNCERTAINTY ESTIMATION FOR TARGET DETECTION SYSTEM DISCRIMINATION AND CONFIDENCE PERFORMANCE METRICS

DISSERTATION

Presented to the Faculty
Graduate School of Engineering and Management
Air Force Institute of Technology
Air University
Air Education and Training Command
in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

David R. Parker, B.S., M.E.S.
Major, USAF

March 2006

AFIT/DS/ENG/06-01

UNCERTAINTY ESTIMATION FOR TARGET DETECTION SYSTEM DISCRIMINATION AND CONFIDENCE PERFORMANCE METRICS

David R. Parker, B.S., M.E.S.
Major, USAF

Approved:

Steven C. Gustafson, Ph.D. (Chairman) — 8 March 2006 — Date

Mark E. Oxley, Ph.D. (Member) — 8 March 2006 — Date

Timothy D. Ross, Ph.D. (Member and Sponsor) — 9 March 2006 — Date

James T. Moore, Ph.D. (Dean's Representative) — 9 March 2006 — Date

Accepted:

M.U. Thomas — 13 March 2006 — Date
Dean, Graduate School of Engineering and Management

Acknowledgements

The results provided in this dissertation would not have been possible if it were not for the dedication of numerous individuals who served many varying, but significant, roles in my program's progression.

Lt Col Goda, my pro-tem advisor, was a tremendous source of consistent, sound guidance and encouragement as I began the program. Lt Col Goda was instrumental in ensuring that my program got off to an appropriate start prior to the formal transition to my dissertation advisor and committee chairman, Dr. Gustafson. Dr. Gustafson's guidance was absolutely critical to the progression of this research. I have great admiration for Dr. Gustafson's character, technical expertise, and professionalism. I looked forward to our frequent discussions; Dr. Gustafson's advice and critique provided the motivation to set short term goals as stepping stones to longer term results; the body of work presented here that resulted over a period of years is simply the culmination of many short term goals that could be accomplished and/or investigated in a period of a day, a few days, or a few weeks. Dr. Gustafson embraces complex topics while still handling such topics in a manner which ensures that they are not made more complex than absolutely necessary. I believe this approach was critical to the progression of this research, and a mindset that I hope to carry with me in the future. The opportunity to receive guidance from and interact with Dr. Gustafson is likely what I will, in the future, appreciate most, when I think back to this program.

The leadership of my sponsor organization, AFRL/SNA, expressed a strong willingness to support the progression of this program from the start. As I neared the completion of my coursework, I was introduced to Dr. Ross, whose subsequent frequent meetings with Dr. Gustafson and myself were critical to the development and refinement of this worthwhile research topic, as well as to the evaluation of the subsequent research's progression. Dr. Ross's guidance and feedback as research sponsor was significant and is very much appreciated.

The feedback and support of my entire dissertation committee, including, near the conclusion of the process, the Dean's Representative, is extremely appreciated. It was an honor to have each of your participation and involvement. Committee member suggestions, from the period of the prospectus formulation through the final dissertation acceptance, without question significantly improved the quality of the research presented here. Very simply, the document would not be the product it is today without your dedication, guidance, and support.

While our exact technical paths in many cases diverged a few years ago, I sincerely appreciate the support of the students with whom I have had the opportunity to get to know and study with here at AFIT. In addition to the faculty and fellow students, I am also deeply grateful for the support of other AFIT professionals who serve many diverse but critical roles (such as the AFIT/ENG support staff, and the library, public affairs, and computer support staffs); their consistent, friendly professionalism towards students such as myself throughout the entire program is very much appreciated.

I would also like to sincerely thank those who motivated me to enter the program in the first place, particularly my co-workers and supervisors at earlier assignments, and previous academic advisors who provided me the necessary encouragement to begin this worthwhile endeavor.

Foremost, I appreciate the gracious support of my family throughout all phases of this endeavor; their understanding and encouragement throughout this entire program was absolutely necessary to succeed in this program.

David R. Parker

Table of Contents

List of Figures

Abstract

This research uses a Bayesian framework to develop probability densities for target detection system performance metrics. The metrics include the receiver operating characteristic (ROC) curve and the confidence error generation (CEG) curve. The ROC curve is a discrimination metric that quantifies how well a detection system separates targets and non-targets, and the CEG curve indicates how well the detection system estimates its own confidence. The degree of uncertainty in these metrics is a concern that previous research has not adequately addressed. This research formulates probability densities of the metrics and characterizes their uncertainty using confidence bands. Additional statistics are obtained that verify the accuracy of the confidence bands. Methods for the generation and characterization of the probability densities of the metrics are specified and demonstrated, where the initial analysis employs beta densities to model target and non-target samples of detection system output. For given target and non-target data, given functional forms of the data densities (such as beta density forms), and given prior densities of the form parameters, the methods developed here provide exact performance metric probability densities. Computational results compare favorably with existing approaches in cases where they can be applied; in other cases the methods developed here produce results that existing approaches can not address.

UNCERTAINTY ESTIMATION FOR TARGET DETECTION SYSTEM DISCRIMINATION AND CONFIDENCE PERFORMANCE METRICS

1. Introduction

This chapter introduces target detection systems and metrics that characterize their performance, reviews existing research on these metrics, summarizes the contributions of this research, and presents the dissertation organization.

1.1 Target detection systems

Decision systems accept input data and generate decision output(s). Examples arise in artificial intelligence, speech processing systems, medical diagnostic systems, and target detection systems. Typically, decision systems make estimates of decision suitability, but do not declare unequivocally that a particular output or action is proper (see [Ross and Minardi, 2004]). A target detection system under test (SUT), the decision system of interest in this research, attempts to estimate the probability that given input(s) contain a target. The inputs are often images, e.g., from synthetic aperture radar (SAR), although the results of this research extend to other types of inputs. Tanks, improvised explosive devices (IEDs), and vehicles containing explosives are some examples of targets.

Estimates of target probability are referred to as scores (see [Wise *et al.*, 2004]). By selecting a threshold score value, all scores greater than a certain value may be declared targets and all scores less than the selected value may be declared non-targets. Thus, the estimates of probabilities may be transferred from a continuous domain to a hard yes/no binary decision on whether or not the input(s) contains a target. To understand the

usefulness of score values and the related varying thresholds, consider two scenarios, labeled A and B.

In scenario A, SUT A attempts to detect a vehicle containing explosives that is a significant distance (two miles) away from a military checkpoint and to label this vehicle as "target". The outcome of a target declaration for this scenario is the raising of barriers and the temporary isolation of the vehicle at a point one-half mile away from the checkpoint so that if, indeed, the vehicle contains explosives, it will not impact either the checkpoint or other nearby vehicles. Once isolated, a more robust stationary monitoring system is used to examine the vehicle. Here, a threshold which results in a declaration of "target" that stops vehicles with explosives but that also stops many vehicles without explosives may be acceptable. A vehicle without explosives that is inadvertently declared a "target" will not be damaged, but will be delayed momentarily. For this example, a threshold that often generates false alarms in that it declares a vehicle without explosives a "target" is appropriate.

In scenario B, SUT B monitors vehicles that approach a business district. In this case, it is impractical to raise barriers. However, a weapon that destroys the vehicle engine is available. If a declaration of "target" is made, then the weapon will be used, otherwise additional sensors will continue to monitor the vehicle. It is easy to see that the threshold selected in this scenario may need to be much higher than the threshold of scenario A so that few false alarms are generated, even if the scores provided by the SUTs are identical.

1.2 Detection system performance metrics

There are two desirable properties of score output from a SUT. The first property is discrimination. Discrimination refers to the ability of an SUT to classify target events as target labels and non-target events as non-target labels. This capability changes depending on the selection of threshold. For a selected threshold, the probability of improperly declaring a non-target event as a target label is referred to as "false alarm

probability". Similarly, the probability of correctly declaring a target event as a target label will be referred to here as "correct detection probability". The purpose of the adjective "correct" here is to emphasize the usage of "correct detection probability" to describe the probability of correctly declaring a target to be a target. An alternative is for "correct detection probability" to denote the probability of correctly labeling an event regardless of whether the event is target or non-target; this alternative is not used here. Note that "false alarm probability" and "correct detection probability" may be replaced by various synonomous descriptions; see the discussion in Section 2.2. The second property is accuracy or relevance and refers to whether or not the estimates of probability that are provided by a SUT are accurate. Both properties are important for SUTs; methods that assist in evaluation of the performance of SUTs with regard to such properties are now introduced.

If the behavior of an SUT over a varying threshold is known, then the discrimination property can be described by a plot of correct detection probability versus false alarm probability. This plot is called a receiver operating characteristic (ROC) curve. For example, consider signals sent from a transmitter to a receiver. The receiver attempts to distinguish "signal" from "noise". The receiver does not know for a selected time sample whether or not a signal has been sent but does measure the amplitude of a demodulated signal at that time. The receiver must choose some threshold value (e.g. 0.3, 0.5, 0.9, etc.), to declare signal; all values greater than the threshold are declared signal and all values less than the threshold are declared "non-signal". For a particular threshold, there is a correct detection probability: among all signals sent, correct detection probability is the percentage declared as signal. Similarly, for a particular threshold, there is a false alarm probability: among all non-signals sent, false alarm probability is the percentage declared as signal. A particular threshold might result in a high correct detection probability but also a high false alarm probability; selection of a different threshold might reduce false alarm probability but also reduce correct detection probability.

The ROC curve described here is formed by varying a single threshold of score. Figure 1.1 shows a score-threshold ROC curve and its generation from target and non-target probability densities of score (hereafter "probability density" is often simply "density"). It is possible to form ROC curves that use multiple thresholds of score, e.g., target is declared between two thresholds, and non-target is declared otherwise. Such ROC curves may be generated by thresholding the likelihood ratio, which is the ratio of target to non-target probability density, as described in the next chapter.

The ROC curve is useful because it provides a tool to examine the trade-off in correct detection probability and false alarm probability. In particular, the ROC curve assists in understanding the relative impact of accepting a higher or lower false alarm probability.

1.3 Discrimination metrics versus confidence metrics

The ROC curve quantifies the discrimination capability of a SUT; the accuracy (or relevance) of estimates of target probability (such estimates are referred to as scores) is of parallel importance to discrimination. In an ideal SUT, the estimates are without error; that is, every provided score is an accurate indication of the probability of obtaining a target given the score. In actual SUTs, estimates of probability may deviate significantly from actual probability. A system that produces an estimate of probability which is very accurate is one that maintains a high degree of "confidence" in results. Thus, the term "confidence" is used to describe the relation of an actual SUT to an ideal SUT in accuracy (or relevance)." Just as the ROC curve characterizes discrimination, the performance of a SUT over all scores can be characterized by a plot of the probability of obtaining a target given a particular score versus score. This plot is called a confidence error generation (CEG) curve.

Both the ROC curve and CEG curve are useful tools for comparing SUTs, thereby determining which SUTs are most appropriate for a particular purpose. Similarly, both the ROC curve and CEG curve may be evaluated for a single SUT to determine whether

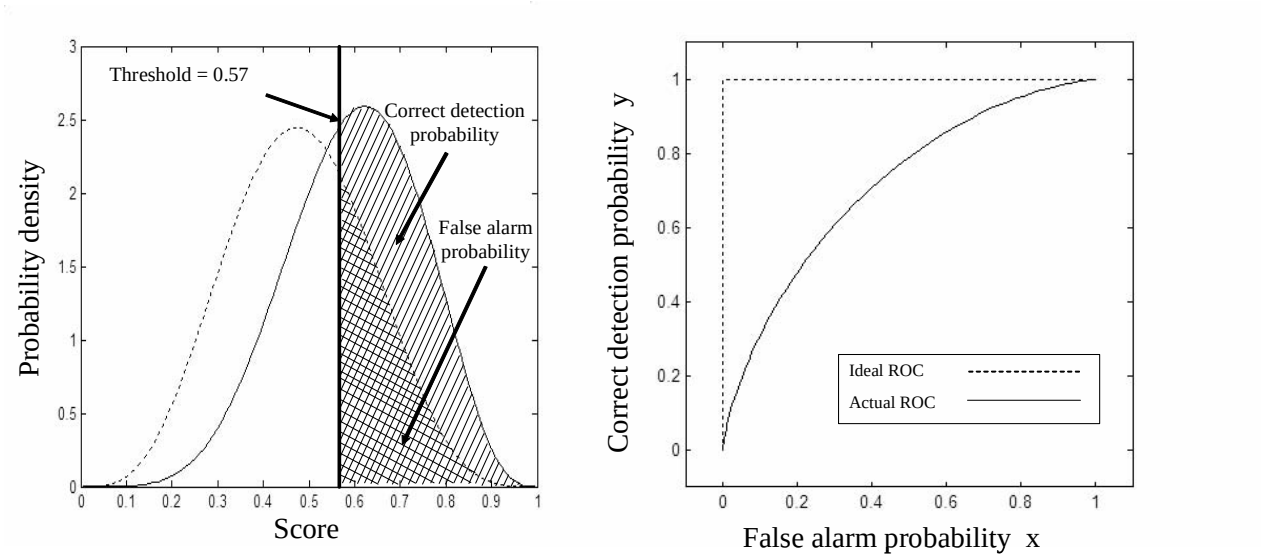


Figure 1.1 Target and non-target densities and the ROC curve performance metric. The ROC curve quantifies the tradeoff in performance between probability of correct detection and probability of false alarm as a decision threshold is changed. The ROC curve has correct detection probability y = 0 for false alarm probability x = 0 and it has y = 1 for x = 1. In the left plot the solid curve is the probability density of target, the dotted curve is the probability density of non-target, and both densities are functions of score. To obtain a score-threshold ROC curve, a threshold is swept across the domain of possible scores from a SUT. For example, at a selected threshold score 0.57, every score greater than 0.57 is regarded as target, and every score less than 0.57 is regarded as non-target. Increasing the threshold leads to a reduced false alarm probability and also a reduced correct detection probability.

or not the SUT is appropriate for a particular purpose. A system that evaluates an SUT through the use of performance metrics (such as ROC curves and CEG curves) is referred to as a Performance Evaluation System. Note that the term metric here refers to a description or characterization of performance or efficiency; this meaning is consistent with recent use of the term metric for software development [Thing, 2002] and is also consistent with recent use of the term metric specific to detection system performance evaluation. For example, the objective of a recent workshop sponsored by the Defense Advanced Research Projects Agency (DARPA), National Institute of Standards and Technology (NIST), and the Institute of Electrical and Electronics Engineers (IEEE), was to define measures and methodologies for evaluating the performance of intelligent systems, and it was entitled "Performance Metrics for Intelligent Systems Workshop" [Messina and Meystel, 2004]. But, mathematically, the term metric is a real-valued function defined on a pair of objects, with specific properties. We apply the formal usage of this term; the entire ROC curve and CEG curve are single comparable descriptions of the overall performance capability of a SUT. Note that the terms "measure" [Ross and Minardi, 2004] and "quantifier" [Schubert *et al.*, 2005] could also be appropriate.

1.4 Evaluation of a system under test

Figure 1.2 shows the relation of the SUT, performance evaluation system, performance metrics (such as the ROC curve and CEG curve), test image inputs, and truth data. To appropriately develop the ROC curve and CEG curve and thus characterize SUT usefulness for a particular purpose, large amounts of test data are desired, where for this data the true state (target or non-target) of the output scores is known. However, such large amounts of test data are typically unavailable or are costly or time-consuming to obtain. As a result, the ability to quantify the uncertainty in the ROC curve and CEG curve performance for limited sets of data is important. If such uncertainty estimates are available, then the range of possible values of the curves given large amounts of data is

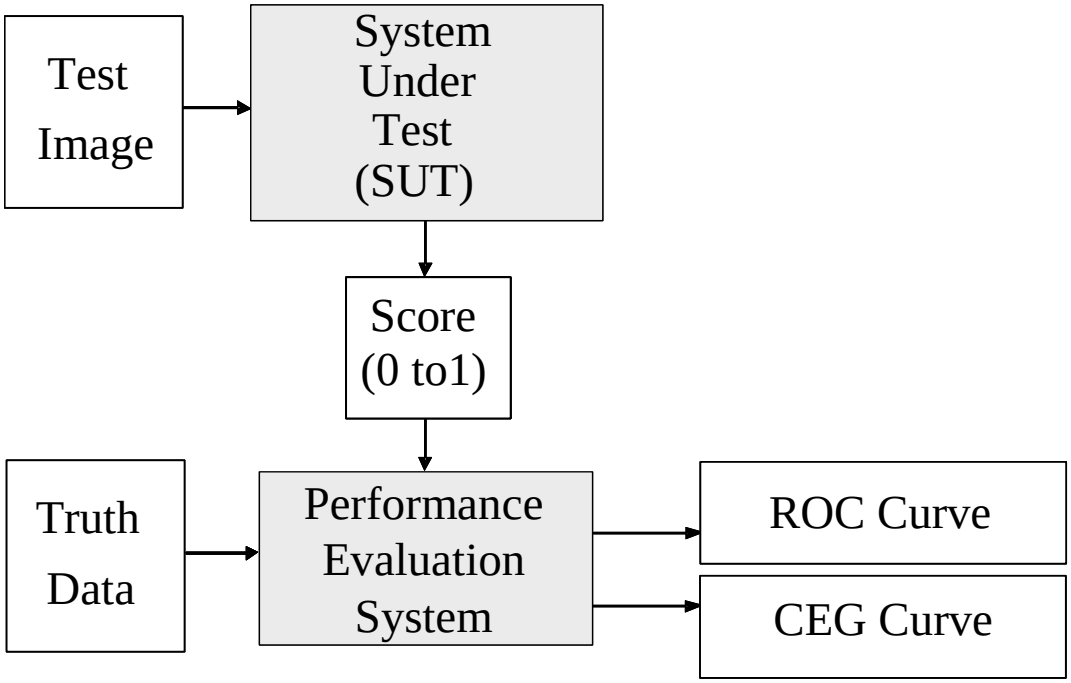


Figure 1.2 Evaluation of a system under test (SUT). The SUT receives a test image and assigns it a score between 0 and 1. A score near one indicates high probability that the test image contains a target, and a score approaching zero indicates a low probability that the test image contains a target. Once the SUT issues a set of scores, a performance evaluation system compares the scores with truth data. The truth data indicates true state of target or non-target in the test image, but does not refer to the entire test image. Performance metrics such as the receiver operating characteristic (ROC) curve and the confidence error generation (CEG) curve are then used to quantify SUT performance.

understood, and acceptable quantification of SUT performance may be possible with limited sets of data. In some cases uncertainty estimates may lead to an informed decision that more data is needed. In other cases, the decision may be that a SUT is suitable for a particular task.

In contrast to current methods for uncertainty estimation, the methods developed here estimate the probability density of a ROC curve based on a Bayesian framework that fully incorporates available information. The Bayesian development incorporates, by definition, all that is known or assumed about the sample score data, the probability density forms for target and non-target scores, and the prior probability densities of the parameters in these forms. For a given set of target samples and non-target samples, assumed sample density models, and prior densities of parameters, there is only one probability density of the ROC curve. The Bayesian formalism permits the generation of this unique ROC curve probability density; descriptive statistics such as the median ROC curve and ROC curve confidence intervals may then be developed, if desired, from this probability density. Non-Bayesian methods either do not fully account for what is known about the data models and prior densities or can only account for this knowledge in an ad hoc manner. The Bayesian probability density of the ROC curve is a full account and is extended in this research to uncertainty estimation of the CEG curve. The results shown here demonstrate improved uncertainty estimation methods for the ROC curve and initiate uncertainty estimation methods for the CEG curve.

1.5 Existing research on performance metric uncertainty

There are existing methods that estimate ROC curve uncertainty. However, these methods typically make unacceptable assumptions; for example, "binormal" methods assume that the target and non-target score densities are either normal or may be made normal after transformation and generally assume that the probability of obtaining a target increases as score increases. "Bootstrap" methods do not make such assumptions

but are inaccurate for relatively small sample size. Still other methods, such as "binomial" methods, may be suitable for estimating the uncertainty in correct detection probability and false alarm probability at a selected threshold but are not appropriate for estimates of the ROC curve over all thresholds.

Figure 1.3 shows a comparison of confidence band results obtained by the method developed here to the most prevalent method in the literature [Metz *et al.*, 1998]. The solid curve in the figure shows the true ROC curve, which is deterministic because it is generated by the target and non-target densities from which the score samples are drawn. A 95% confidence band based on an observed set of 30 target and 30 non-target score samples is shown for the method developed here. The Metz method (which is a binormal approach) produces a 95% confidence band that is wider and therefore less informative than the band for the method developed here, assuming that the method developed here is accurate with respect to the assumed density forms and the prior densities of parameters. Chapter 3 considers the analytical justification for the method developed here, and Chapters 4 and 5 demonstrate its accuracy. Chapter 2 examines the Metz method and other ROC curve confidence interval methods in detail. The method developed here performs favorably in comparison with the other methods (where suitable comparison is possible). More importantly, the method developed here shows the viability of a flexible Bayesian framework and enables the development of alternative descriptive statistics (such as initially considered in [Parker *et al*, 2005a, 2005b]). The method is directly applicable to other metrics, such as the CEG curve (the CEG curve is detailed in Section 2.3; see also [Parker *et al.*, 2005c]). This framework permits changes in model assumptions; the Metz method and most other approaches do not allow such changes.

1.6 Summary of contributions of this research

The research reported here uses a Bayesian framework to characterize the uncertainty of target detection performance metrics. The result is an improved understanding and

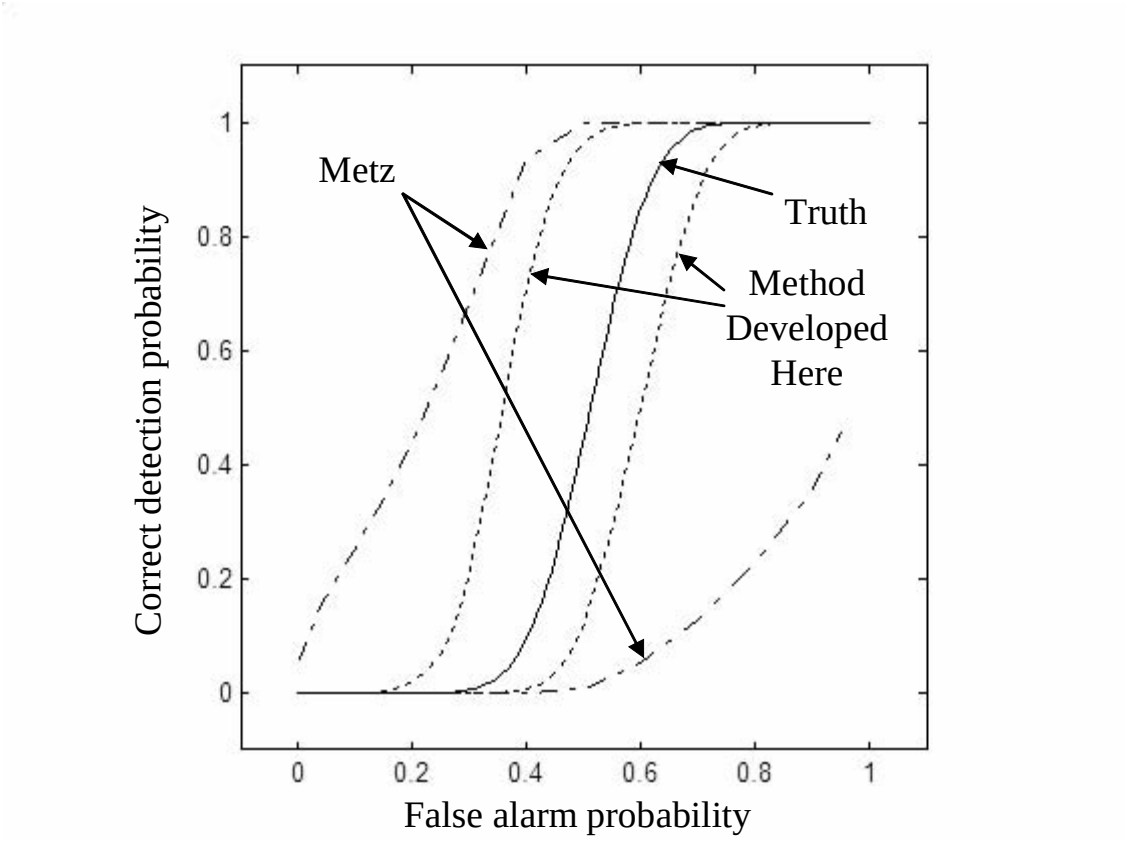


Figure 1.3 Comparison of method developed here with the method of Metz [Metz *et al.*, 1998]. Here 30 target and 30 non-target samples are drawn from beta densities with target mean 0.715, target standard deviation 0.01, non-target mean 0.715, and non-target standard deviation 0.046; the solid line is the true ROC curve. Note that the Metz ROC curve 95% confidence band is extremely wide (and uninformative) compared to the 95% confidence band obtained using the method developed here. The software package "ROCKIT" is used to generate the confidence intervals for the method of Metz.

quantification of ROC curve and CEG curve uncertainty for target detection applications. The framework develops ROC or CEG curve probability densities, which completely describe curve uncertainty for given samples of target and non-target scores, assumed density forms for these scores, and assumed prior densities of the parameters that specify these forms. From the ROC or CEG curve densities, a transition to descriptive statistics, such as median curves or 90% confidence intervals, is made. The framework is fully Bayesian and for the given samples, density forms, and prior parameter densities it provides exact performance metric probability densities.

The framework is also numerically tractable, and the calculated ROC and CEG curve densities yield substantial improvements over existing ROC curve uncertainty estimation methods. These improvements are emphasized qualitatively in the identification of fundamental weaknesses inherent in existing ROC curve uncertainty estimation methods; in addition, quantitative comparisons are made which verify that the approach developed here compares favorably with previous approaches. Further, the uncertainty estimation process is shown to seamlessly transition to the CEG curve, a metric that previously has been of limited use due to a lack of appropriate methods for estimating its uncertainty, especially for limited amounts of data. From the framework developed here, CEG curve uncertainty estimates can now be robustly understood and obtained even for low numbers of samples. Thus, for the CEG curve, the research presented here formulates a robust method for uncertainty estimation where alternatives do not exist; for the ROC curve the research presented here offers a significantly improved method of uncertainty estimation for which the alternatives are limited by inability to handle low numbers of samples and/or by restrictive model assumptions.

1.7 Organization of this dissertation

Chapter 2 provides background on the uncertainty estimation problem considered here, provides a review of the literature, and identifies weaknesses in existing uncertainty

estimation methods for ROC and CEG curves. Chapter 3 describes and develops both analytical expressions and numerical approximations for the ROC curve probability density. This ROC curve density is then used in Chapter 4 to obtain and verify descriptive statistics, such as median ROC curves and 90% confidence intervals. Chapter 5 provides quantitative comparisons of the method developed here to previous methods of confidence interval estimation. Chapter 6 summarizes accomplishments and contributions and identifies areas of interest for future work.

2. Background

Performance metrics such as the ROC curve quantify the capability of a target detection system to distinguish between target and non-target inputs. Other performance metrics such as the CEG curve examine the relevance of the detection system outputs. This research develops improved methods for estimating the uncertainty of these metrics and many other types of target detection performance metrics. Since a ROC curve for one target detection system under test (SUT) may be compared to a ROC curve for a second SUT, the ROC curve and CEG curves are referred to here as metrics, although the curves are not scalar values.

As discussed in the introduction, the outputs of a target detection SUT are typically estimates of the probability of target. Such estimates are referred to as posterior probability estimates and are critical in appropriate decision making (see [Bishop, 1995]). For example, the speech processing community often makes estimates through the use of cross-entropy (see relation of speech processing techniques to the target detection field by [Ross and Minardi, 2004]). A speech processor typically examines a portion of observed input speech and attempts to match this input with plausible phonetic sounds. The processor does not declare with absolute certainty that a portion of observed speech is a particular sound; however, it estimates the probability of a sound or group of sounds. Then, when groups of adjacent input speech are examined, the estimates of probability are used to formulate words, phrases, and sentences. Similar to the speech processor example, a SUT does not declare a target with certainty but instead estimates the probability that given input(s) contain a target.

Specific to the focus here on target detection, development and use of the CEG curve performance metric by the Sensors Directorate of the Air Force Research Laboratory motivates this research (see [Ross and Minardi, 2004] and [Wise *et al.*, 2004]) in that CEG curve uncertainty was not well characterized. Thus, the methods developed here

are first applied to the ROC curve and are then are used to estimate CEG curve uncertainty. However, existing approaches to ROC curve estimation are inadequate, particularly when low numbers of inputs are available and when normality assumptions are invalid; these conditions are both common constraints for target detection systems. The methods developed here show improved results compared to existing ROC curve uncertainty estimation approaches and are, in fact, optimal (see Section 2.5 and method development in Chapter 3). The results of this research also benefit the wide range of fields that use ROC curves (e.g., medical decision making, machine learning).

2.1 Target detection systems and their performance evaluation

Figure 1.2 in Chapter 1 shows the relation between a test input, the SUT, the performance evaluation system, and performance metrics. (Note that although a test image is used as an example, the process also directly applies to other types of test inputs.) For each test image the SUT outputs a score between zero and one. This score provides an estimate of the probability that the image contains a target. A score of one estimates a probability of one that the image contains at least one target, and a score of zero estimates a probability of one that the image does not contain a target. The performance evaluation system knows truth for test cases; that is, whether an image actually contains a target or not. The performance evaluation system has two input types, the scores for many images from the SUT and the truth (target or non-target) associated with each image. Performance metrics such as the ROC curve and the CEG curve are outputs of the performance evaluation system. The area under the ROC curve (AUC) value and the CEG curve summary metric of root square deviation (RSD) value are also considered. A key distinction is that the ROC curve and AUC value describe how well a system is able to discriminate between target and non-target without regard to whether or not the scores are accurate estimates of the probability of target, whereas the CEG curve and RSD value are metrics that describe such accuracy (or relevance) [Ross and Minardi, 2004].

2.2 ROC curves and AUC values

The ROC curve (see [Lusted, 1971] and [Swets, 1988]) is a plot of probability of correct detection versus probability of false alarm based on a varying threshold for detection. Figure 1.1 of Chapter 1 shows such a plot; this figure also demonstrates the calculation of probability of correct detection and probability of false alarm for a single selected threshold. The ROC curve quantifies the trade-off in performance between correct detection probability (y) and false alarm probability (x) as a decision threshold (t) is changed (see [Alsing, 2000]). The ROC curve derives its name, *receiver* operating characteristic, from its original application, which focused on radio applications [Wickens, 2002]. Beginning with its original application in the 1950s, it has been used in many other applications, such as the target detection performance metric that is the focus of this research, medical decision making (e.g. quantifying the probability of a disease occurring given a biological marker; see [Hanley, 1999]), and machine learning (see [Macskassy and Provost, 2004]).

Three formal definitions related to the ROC curve are as follows.

(1) Let E be the population set of test images, where the test images either contain a target (target images) or do not contain a target (non-target images). Based on an estimate of whether each image $\varepsilon \in E$ actually has a target, an SUT produces a data score d, where $d \in D = [0, 1]$. Thus, the SUT maps E to D denoted by $E \overset{SUT}{\rightarrow} D$. Let $\Theta = [0, 1]$. For each $\theta \in \Theta$, let a_θ be a classifier mapping D into a label set L denoted by $D \overset{a_\theta}{\rightarrow} L$ where $L =$ {target declaration, non-target declaration}. Thus, the classifier system is $E \overset{SUT}{\rightarrow} D \overset{a_\theta}{\rightarrow} L$. For any element $\varepsilon \in E$, $d \in D$, and $l \in L$, choice of θ specifies the classifier, and Equation (2.1) specifies the label for the score-threshold method:

$$l = \left\{ \begin{array}{c} \text{target declaration: } d >= \theta \\ \text{non-target declaration: } d < \theta \end{array} \right\}. \tag{2.1}$$

The *threshold for detection* is t,where t is a specified θ. The above is adapted from Schubert, Oxley, Bauer [Schubert *et al.*, 2005], who provide a similar classifier definition but with application to a more general classifier system, rather than the score-threshold application of interest here.

(2) Let E_{target} be the subset of all $\varepsilon \in E$ that contain target images. Let $D_{\text{target}} \subset D$ be the subset of all $d \in D$ corresponding with E_{target}. Let $s \in (\text{-}\infty, \infty)$. Let $g(s)$ be the target score probability density formed by all D_{target}, where s is a scalar random variable. The *correct detection probability* is

$$x = \widehat{G}(t) = \int_t^{\infty} g(s)ds. \tag{2.2}$$

(3) Let $E_{\text{non-target}}$ be the subset of all $\varepsilon \in E$ that contain non-target images. Let $D_{\text{non-target}} \subset D$ be the subset of all $d \in D$ for $E_{\text{non-target}}$. Let $s \in (\text{-}\infty, \infty)$. Let $f(s)$ be the non-target score probability density formed by all $D_{\text{non-target}}$. Specify $t \in (\text{-}\infty, \infty)$. For the score-threshold method described by Equation (2.1), let $t = \theta$. The *false alarm probability* is

$$y = \widehat{F}(t) = \int_t^{\infty} f(s)ds. \tag{2.3}$$

Typically a threshold for detection (or simply, threshold) is applied either to score or likelihood ratio, where the likelihood ratio is the target probability density divided by the non-target probability density. The threshold of interest here and described in the above definitions is score-threshold (as described in Equation (2.1)), because the primary objective is to use ROC curves and AUC values (and other performance metrics) to quantify whether a SUT is performing optimally, rather than to use the ROC curves and AUC values to optimize SUT performance. If the threshold for detection is set at zero (i.e., all score values are declared as targets), 100% of targets are detected, but this choice also results in a probability of false alarm equal to one. If the threshold for detection is set at one, no false alarms occur, but the probability of correct detection is zero. An ideal

ROC curve has a correct detection probability that equals one for all false alarm probability greater than zero. Thus, an ideal ROC curve has an AUC value that equals one, whereas a non-discriminating ROC curve has an AUC value that equals 0.5. The AUC value is the integral from 0 to 1 of correct detection probability y as a function of false alarm probability x. The ROC curve is the set $\{(x, y) \in [0, 1]\text{x}[0, 1] | y = r(x) \forall x \in [0, 1]\}$. If r is the function that generates the ROC curve, so that $y = r(x)$, then

$$AUC(r) = \int_0^1 r(x)dx. \tag{2.4}$$

The research here focuses on this score-threshold ROC curve. However, an alternative method, which is not desirable for comparison of multiple SUTs by an evaluator (assuming that the evaluator only has access to scores provided by the SUTs), but that can be a desirable tool for SUT improvements, uses maximum likelihood (via the Neyman-Pearson Lemma; see [Scharf, 1991]) to develop the ROC curve. A likelihood-ratio-threshold ROC curve (see [VanTrees, 1968] and [Scharf, 1991]), is generated by thresholding the ratio of the target and non-target densities, and is consequently convex (this curve has a negative second derivative at each false alarm probability). A likelihood-ratio-threshold ROC curve allows multiple positive (i.e., target) decision regions across the range of possible score values, whereas a score-threshold ROC curve allows only one positive decision region (see [VanTrees, 1968], [Shanmugan and Breipohl, 1988], [Barkat, 1991], and [Scharf, 1991]). Figure 2.1 compares the procedures for generating a score-threshold ROC curve and a likelihood-based ROC curve. The score-threshold ROC curve always has an AUC value equal to or less than the likelihood-ratio-threshold ROC curve, assuming that the likelihoods are accurately known when designing the detection system. Note that while the target and non-target densities are of beta density form in the example used in the

figure, this property holds for any probability density (e.g. Gaussian, beta, mixture of beta, etc.).

To understand the rationale for using score threshold, consider a system under test that declares a score of "0" for all targets and a score of "1" for all non-targets. Since the scores provided by a SUT are estimates of the probability that the evaluated image is a target, this performance is obviously poor. The corresponding score-threshold ROC curve has an AUC of zero, affirming that the system is performing poorly. In contrast, a likelihood-ratio-threshold-ROC curve estimated ROC has an AUC of one. Thus a likelihood-ratio-threshold ROC curve may be of significant interest for developing a target detection system, but a score-threshold ROC curve is most relevant to the objective of evaluating system performance.

Figure 2.1 shows deterministic target and non-target densities, each for two specified parameters (see Equation (3.1)) and compares a score-threshold approach with a likelihood-ratio-threshold approach. Note that while beta densities are the focus of these figures, the methods developed here extend to other density forms (see Figure 3.4 and related discussion in Section 3.1).

A theorem that provides an analytical form for the ROC curve is as follows.

Theorem 2.1 Score-threshold ROC curve

Let $f(s;u)$ and $g(s;v)$ be densities of s given parameters u and v, where s is a real-valued random variable between zero and one, $s \in [0,1]$, $f(s;u)$ is the non-target score probability density, $g(s;v)$ is the target score probability density, u is a parameter vector that specifies the non-target score density, and v is a parameter vector that specifies the target score density. Let f and g be integrable over $[0,1]$ for each u and v, and for each $t \in [0,1]$ define

$$\widehat{F}(t;u) = \int_t^1 f(s;u)ds = x, \tag{2.5}$$

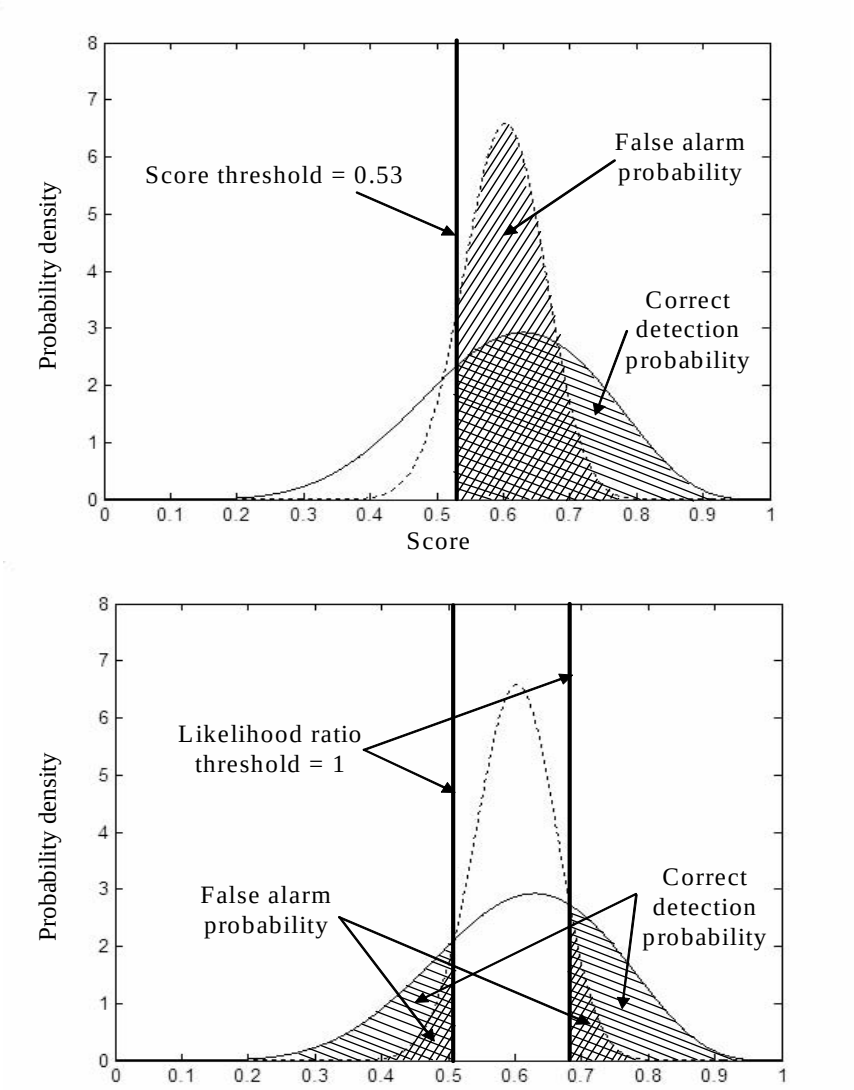


Figure 2.1 Comparison of score-based and likelihood-based ROC curve generation. In the score-based threshold approach (top figure), a probability of correct detection is calculated by selecting a threshold for score (e.g., 0.53), and integrating over the target density (solid curve) from that threshold to 1. Similarly, a probability of false alarm is calculated by integrating the non-target density (dotted curve) over the same domain. The values for probability of correct detection and probability of false alarm form a point on the ROC curve, and the ROC curve is formed by varying the threshold from 0 to 1. In the likelihood-based approach (bottom figure) the likelihood ratio, which is the ratio of target to non-target densities, is thresholded (e.g., at 1), so that in general there is more than one correct detection and false alarm region.

and

$$\widehat{G}(t;v) = \int_t^1 g(s;v)ds = y, \tag{2.6}$$

so that

$$\widehat{F}(t;u) = 1 - F(t;u), \tag{2.7}$$

and

$$\widehat{G}(t;u) = 1 - G(t;u), \tag{2.8}$$

where $F(t;u)$ and $G(t;v)$ are cumulative probability distributions.
If the inverse of $\widehat{F}$ exists for every u, then the score-threshold ROC curve is (by implicit and inverse function theorems; see [Olmstead, 1961])

$$y = r(x;u,v), \tag{2.9}$$

where

$$r(x;u,v) = \widehat{G}(\widehat{F}^{-1}(x;u);v). \tag{2.10}$$

Equivalently, $y = r(x;w)$ and $r(x;w) = \widehat{G} \circ \widehat{F}^{-1}(x;w)$, where w concatenates u and v (i.e., $w = [u_1\ u_2\ ...\ v_1\ v_2\ ...]$).

The proof is in Appendix A-1.

Note that if u and v are fixed, they may be removed in the above formulas (e.g., for fixed u, $f(s) = f(s;u)$); however, retaining u and v is important in later ROC curve density development where the parameters are not fixed. The parameters u and v (or w) characterize the target and non-target densities of score; the Bayesian approach does not require the assumption that such parameters are stochastic (see [Gregory, 2005] and [MacKay, 2003]), but it is acceptable to handle them as random variables (see [Schervish, 1995]). However, it is common practice to simply refer to u and v $(or\ w)$ as

parameters (see [Schmitt, 1969] and [Kass and Raftery, 1995]) or "random parameters" (see [Robert, 2001]). Here the term "parameters" is used.

The AUC value integrates the area under the formed ROC curve; an ideal AUC value is one (see Equation (2.4)). A large AUC value (an AUC value near one) is due to sufficient separation between the target and non-target densities, rather than whether or not the score values are appropriate estimates of the probability of a target. An analysis of AUC value applicability in evaluating pattern recognition systems is given by Alsing [Alsing, 2000], and additional analysis specific to AUC value applicability is provided by Bradley [Bradley, 1997]. Note that the AUC value is a number, but the ROC curve is a function. Thus, the ROC curve is a performance metric that generates one AUC value, but a given AUC value may be generated by many different ROC curves. If most of the target density is greater than some score and if most of the non-target density is less than this score, then the AUC is close to one. In this situation, the ROC curve does not indicate whether or not the scores are appropriate estimates of the probability that the observed image is a target, but the CEG curve and the RSD value metric provide this indication. For target detection system evaluation, the score-threshold ROC curve plots the probability of false alarm and probability of correct detection values achieved by varying a score threshold. However, this ROC curve does not indicate the threshold that is required to obtain a particular probability of false alarm and probability of detection. For some applications, it is of interest to examine only particular regions of the ROC curve; for example, in cases where a false alarm probability greater than a certain value is not relevant.

Correct detection probability is used here to refer to the probability of correctly declaring a target to be a target. False alarm probability is used here to refer to the probability of incorrectly declaring a non-target to be a target. The terms referred to here as correct detection probability and false alarm probability also have other designations. The use of the term correct detection probability here can be replaced by "detection probability" or

"true positive probability". Similarly, "false alarm probability" may be replaced by "false positive probability (see [Hill *et al.*, 2003]). In medical research, "specificity" and "sensitivity" are often used instead of "correct detection probability" and "false alarm probability"; correct detection probability as used here can be substituted for sensitivity, and false alarm probability can be substituted for one minus specificity. The use of correct detection probability here reinforces its usage in "correctly" declaring a target to be a target.

Many radar applications [Hall *et al.*, 1991] focus on low false alarm probabilities, e.g., probabilities on the order of 10^{-14} to 10^{-2} may be appropriate [Raemer, 1997]. In such applications, estimating the uncertainty of the full ROC curve may seem to be of limited practical interest. However, the success of these applications depend on detection system performance. Chapter 1 discussed the practical importance of both low and high false alarm probability in specific examples, and interest in the full range of false alarm probabilities is consistent with recent target detection focused research (e.g., [Zelnio *et al.*, 2005]). As an additional example, consider an unmanned aerial vehicle (UAV), such as the Global Hawk Unmanned Aerial Reconnaissance System. Global Hawk flies at an altitude of 65,000 feet, and has two synthetic aperture radar modes: wide area search mode (1.0 meter resolution) and spot image mode (0.3 meter resolution) [Curiel, 2005]. The wide area search mode can cover a wider area in a fixed amount of time than the spot mode (40,000 square miles versus 3,000 square miles in 24 hours [Humphlett, 2004]), but the wide area search mode has lower resolution [Humphlett, 2004] [Curiel, 2005]. Thus, Global Hawk may declare objects to be targets of interest in the wide area search mode with high false alarm probability permitted, and it then may use the declarations to subsequently examine the objects more closely in spot mode. Note that even in spot mode, a high false alarm probability may be acceptable if the outcome of a target declaration results in a closer examination by a lower flying air-based or ground-based detection system. Finally, note that even for radar systems with very low false alarm probability requirements, accurate performance at higher false

alarm probabilities may be important for monitoring proper system function [Hall *et al.*, 1991]. The methods developed in Chapters 3 through 5 are applicable to the full range of false alarm probability.

2.3 CEG curves and RSD values

The CEG curve describes the accuracy (or relevance) of the target/non-target score values, that is, the curve describes whether the target/non-target score values are appropriate estimates of the actual probability of observing a target. In contrast, the ROC curve describes how well the target and non-target scores are separated [Wise *et al.*, 2004]. Recall that a SUT outputs both target and non-target scores, and if the scores are accurate, then the probability of target given score equals the assigned score; that is, if an ideal SUT generates 100 scores of 0.6, then 60 of these scores are targets and 40 are non-targets. Here, "ideal" refers to an SUT that generates scores (estimates of probability of observing a target) which always equal the true probability of observing a target given the score.

The *RSD value* is defined as

$$RSD = \sqrt{\int_0^1 (P(T|s) - s)^2 p(s) ds}, \tag{2.11}$$

where, using Bayes' rule,

$$P(T|s) = \frac{g(s|T)P(T)}{g(s|T)P(T) + f(s|N)P(N)}, \tag{2.12}$$

and s is a scalar random variable between zero and one, $s \in [0, 1]$, $P(T|s)$ the probability of target event given score, $g(s|T)$ is the density of score given target event, $f(s|N)$ is the probability density of score given non-target event, $p(s)$ is the prior probability density of the score (without regard to target or non-target), $P(T)$ is the prior probability of target

event, and $P(N)$ is the prior probability of non-target event ($P(N) = 1 - P(T)$). The *CEG curve* is defined as a plot of $P(T|s)$ versus score. Similar to the relation of AUC to ROC, whereas RSD is a value, $P(T|s)$ is a function, and the curve that it forms as score varies between zero and one is the CEG curve as shown in Figure 2.2.

Note that many distinct target and non-target densities result in ROC curves that are close to an ideal AUC value of 1. For example, choose any target beta density mean and non-target beta density mean. If the target density standard deviation is sufficiently small and if the target density mean is greater than the non-target density mean, the AUC value approaches one. For the RSD value, only more specific special cases of target and non-target densities approach the ideal RSD value of zero. These special cases include: (a) target density approaches an impulse function (i.e., a Dirac delta function density or distribution) at a score of 1 and the non-target density approaches an impulse function at a score of 0 and (b) target density and non-target densities approach impulse functions at a score of 0.5, and (c) the ratio of the target density to the non-target density is equal to the value of score for all scores.

Figure 2.3 illustrates the process that forms a CEG curve. The lower two plots compare the RSD value described by Equation (2.11) with an unweighted RSD, which does not depend on overall density of score. The weighted RSD value used here is generally preferable (see [Parker *et al.*, 2005c]), because scores that occur infrequently do not increase the RSD value in the weighted method. Figure 2.4 shows similar plots, but the target and non-target densities in this figure generate a more ideal CEG curve and a lower RSD value.

2.4 Relation of performance metrics to SUT evaluation

The objective here is improved evaluation of SUT performance and in particular on improving the ability to describe uncertainty in performance. However, first consider the case where the scores that a SUT outputs for a population set of target and non-target

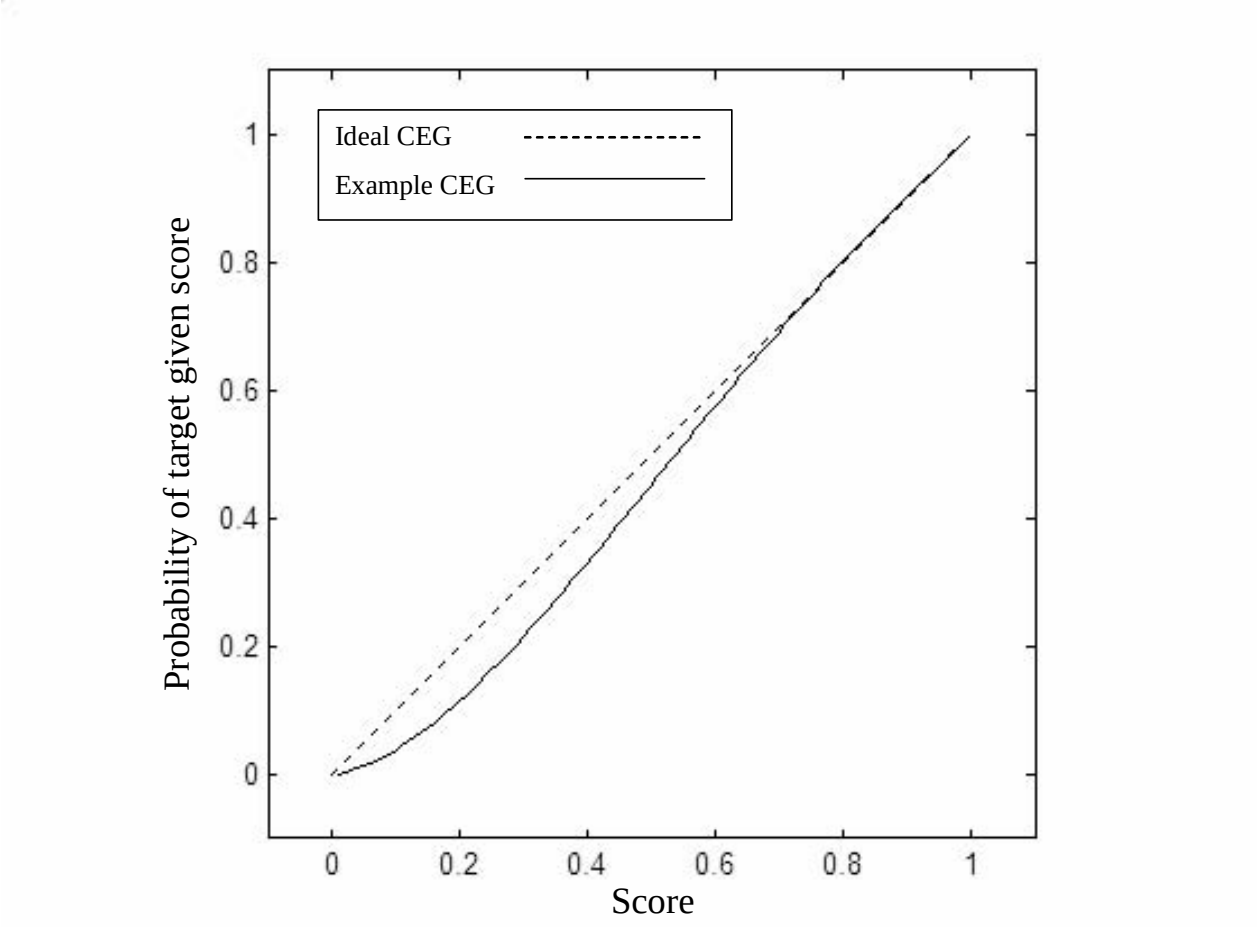
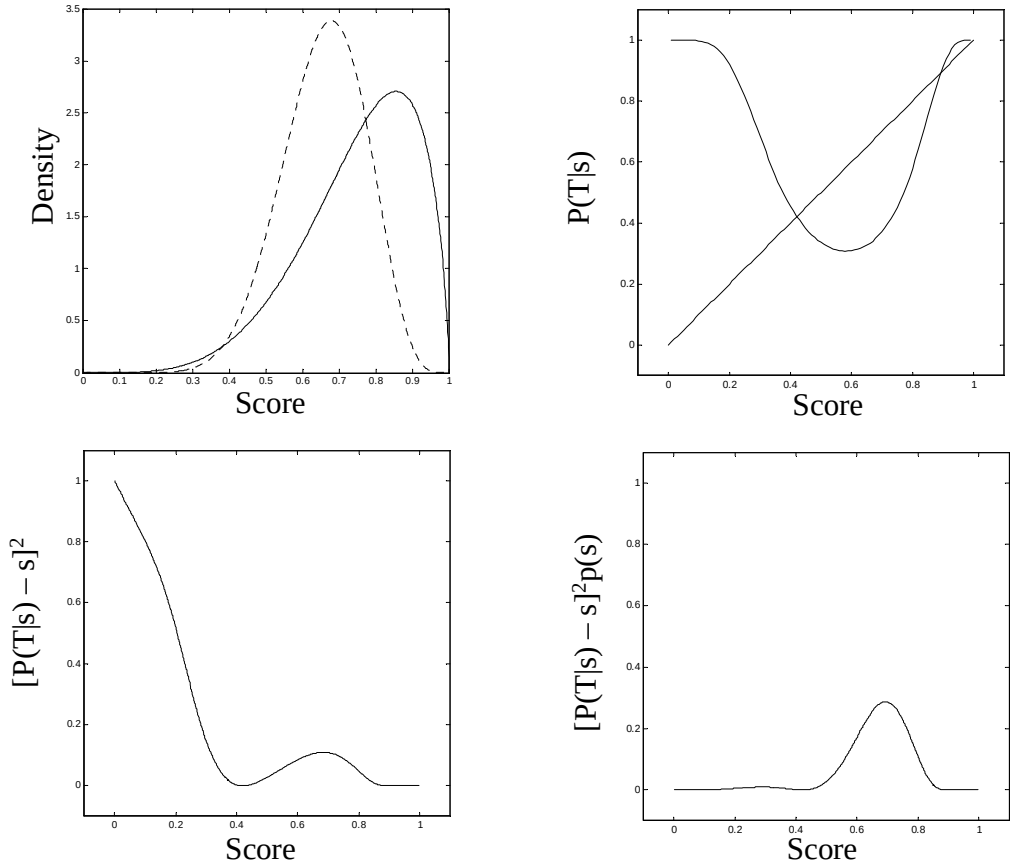
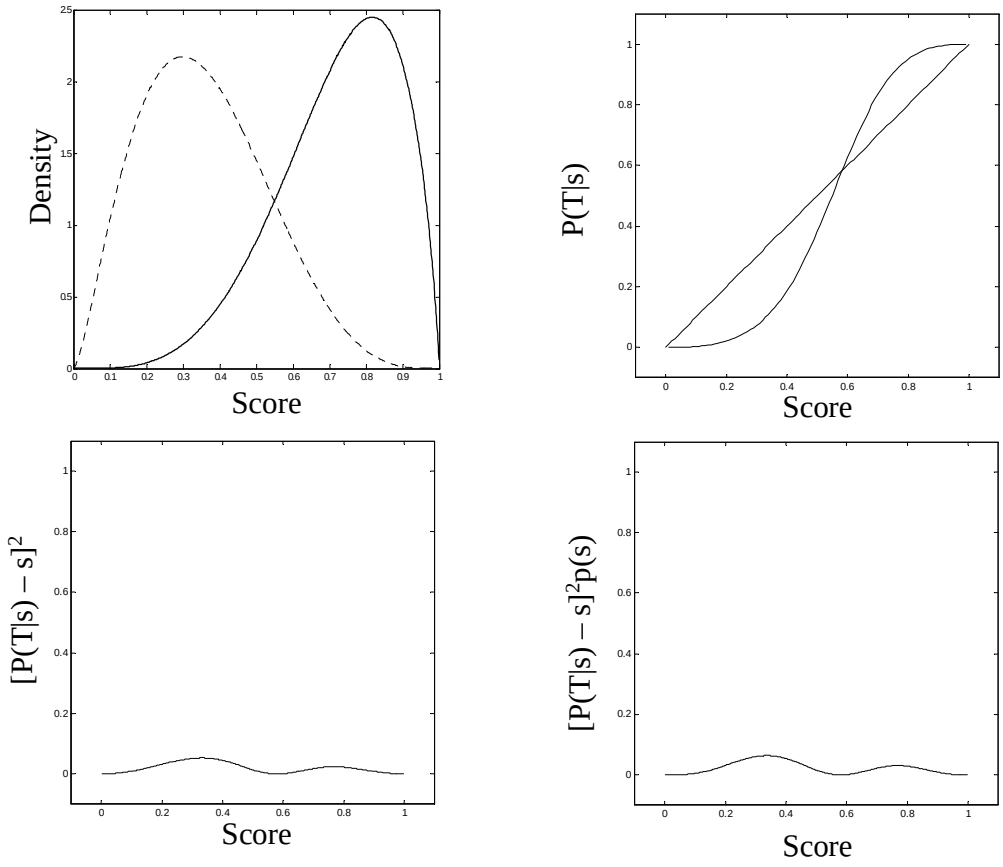


Figure 2.2 The CEG curve. The CEG curve describes the relevance of scores produced by a SUT. For example, if an ideal SUT produces 100 scores at values near 0.75, 75 of the scores are targets and 25 are non-targets. The RSD value summarizes the CEG curve metric and is the root-mean-squared difference of the probability of target given score and score weighted by the density of score. The ideal CEG curve is the dotted 45 degree line; an actual CEG curve is shown by the solid line. At its tails, the density of score may approach zero, yet the deviation of P(T|s) from ideal at these tails may be significant. Therefore the incorporation of the density of score as a weight is important.



$$RSD_{unweighted} = \sqrt{\int_0^1 [P(T|s) - s]^2 ds} = 0.4725 \qquad RSD = \sqrt{\int_0^1 [P(T|s) - s]^2 p(s) ds} = 0.2481$$

Figure 2.3 Target and non-target densities, CEG curves, and RSD values. As shown in the equations on the plot, a RSD value can be weighted or unweighted. The weighted RSD value is affected by the overall densities of score. The top left plot shows a target density (solid line) and non-target density (dashed line). The top right plot shows the CEG curve as the probability P(T|s) of a target versus score. The bottom two plots show the quantities that are integrated to obtain unweighted or weighted RSD value. In an ideal SUT, P(T|s) follows the 45 degree line shown in the top right figure.



$$RSD_{unweighted} = \sqrt{\int_0^1 [P(T|s) - s]^2 ds} = 0.1423 \qquad RSD = \sqrt{\int_0^1 [P(T|s) - s]^2 p(s) ds} = 0.1512$$

Figure 2.4 As in Figure 2.3, but for different target and non-target densities. Here the unweighted RSD value and the (weighted) RSD value are approximately equal because the regions where P(T|s) deviates greatly from score (for example, scores between 0.01 and 0.3) also have high overall score density.

scores are known. In this case, the exact ROC curve and exact CEG curve may be calculated, as described in Section 2.2. The exact ROC curve presents the full set of possible correct detection and false alarm probabilities, and this set indicates the capability of a SUT to differentiate between target and non-target scores. The score-threshold ROC curve that is the focus here (rather than a likelihood-ratio-threshold ROC curve) provides the additional indication, through curve shape, of whether or not the SUT produces appropriate output. For an ideal SUT, an increase in score makes it increasingly likely that a target is observed. A score-threshold ROC curve reveals this result; however, a likelihood-ratio-threshold ROC curve assumes, but does not indicate, this behavior. A likelihood-ratio-threshold ROC curve is always convex; a score-threshold ROC curve is only convex when an increase in score increases the probability of observing a target for all scores. The exact CEG curve describes whether or not the scores provided by a SUT are relevant; that is, whether or not the scores that an SUT generates are representative of the actual probability of target given score. Further, the combined examination of the ROC curve and CEG curve characteristics of a SUT provide robust tools for comparing one SUT with another. The related summary AUC and CEG values also provide useful tools for comparison; however, the curves themselves enable particular probability of false alarm regions (in the case of the ROC curve) and particular score regions (in the case of the CEG curve) to be isolated and analyzed.

A key motivation for this research follows from the fact that in practice, there is only a finite, and often small, set of score samples available to a target detection system. There are methods to estimate the ROC curve and CEG curve from such sets; however, understanding the uncertainty in the estimates may be more important, particularly for low numbers of score samples, than estimating the most likely ROC and CEG curves (such maximum-likelihood estimates are inherently inaccurate for low numbers of samples; see the discussion in Section 3.1).

This research focuses on improved methods of estimating uncertainty in ROC curves, and then extends the development to CEG curves. As discussed in the literature review of Section 2.7, current methods of ROC curve uncertainty estimation make unacceptable assumptions or are only appropriate when sample size is very large, and thus existing methods are not suitable to extend to the CEG curve.

The ROC curve uncertainty estimation methods developed here provide results that can be compared with results in the existing literature and that can then be extended to the CEG curve uncertainty estimation problem. The techniques developed here are unprecedented in ROC curve uncertainty estimation (see the literature review of Section 2.7 and related quantitative ROC curve confidence interval comparisons of Chapter 5). Further, the ROC curve confidence interval framework makes flexible assumptions; even when quantitative comparisons with other methods appear somewhat comparable, these methods generally have unacceptable weaknesses (see Chapters 2 and 5). Note that if only a "best estimate" of the ROC curve is required, there are suitable alternatives to the method developed here (e.g. maximum likelihood), particularly when the prior probability densities are diffuse. While the ROC curve and CEG curve are estimated by the method developed here, obtaining these curves is not the primary motivation. The method developed here focuses on uncertainty estimation, and the primary description for such uncertainty estimation here (and in the literature) is confidence intervals. Confidence intervals are important because for the low numbers of samples that are typical for target detection applications, any best (e.g., maximum likelihood) estimate of the ROC curve may not be close to the actual curve. Thus confidence intervals are of practical interest because they provide a description of the range of possible values for a ROC curve if large (approaching infinite) sets of samples were actually tested.

The beta probability density, while possessing many desirable qualities for the methods developed, is only an example. It is the density that has maximum entropy among all densities that are non-zero on a fixed interval, subject to specific constraints (see

[Gokhale, 1975]) that may be related to mean and variance. However, the analytical expressions developed in Chapter 3 are general and may be applied to alternative density models.

2.5 Bayesian probability densities

The methods developed here use a fully Bayesian framework to develop probability densities for ROC curves and other target detection performance metrics. A Bayesian framework incorporates input samples (such as target and non-target samples), model (such as assuming that the samples are modeled with a Gaussian density), model parameters (such as mean and standard deviation), and prior density assumptions (for example, assuming uniform prior probabilities of means from zero to one and standard deviations from zero to two). Then, the Bayesian framework combines such inputs and assumptions and produces a posterior probability density of an output of interest, such as the ROC curve here. Note that the posterior probability density may be updated if more input samples are available, but that this density is the actual, complete solution for the available samples, model, and priors (see [MacKay, 2003] and [Carlin and Louis, 2000]). In developing the posterior probability density (which the Bayesian framework makes possible), the observed data samples are handled as fixed known input observations. In alternative (frequentist-based) approaches, there is an upfront focus on describing the randomness of the data samples (e.g., using probability statements and confidence intervals), thus making estimates about what samples might have been produced if more samples were available. These estimates are then applied to make follow-up statements about the result of interest (the ROC curve and CEG curve in the case of this research). In contrast, in a Bayesian framework it is the evaluated model parameters that are handled as unknown parameters (see discussion in Section 2.2 and [Bolstad, 2004]). Neither of the two methods ignores uncertainty; both frequentist-based and Bayesian methods make attempts to quantify uncertainty. However, a benefit of the Bayesian framework is that it permits the progressive development of a full, complete, posterior

probability density for the result of interest (e.g., development of the posterior probability densities for the ROC curve and CEG curve in the case of this research) prior to the development of further descriptions such as confidence intervals, median estimates, etc. This developed posterior probability density fully describes the uncertainty of the result of interest based on the available observed data samples, and the model, and prior knowledge. Gregory [Gregory, 2005] provides a detailed discussion and further comparison of frequentist-based and Bayesian approaches. A similar framework was developed in the early 1990s for neural network applications (see [MacKay, 1992a, 1992b] and [Bishop, 1995]); however it has not previously been applied to target detection performance metrics. The densities developed using the framework are characterized here by descriptive statistics, such as median estimates, confidence intervals for ROC curves, and also by statistics that characterize the accuracy of the confidence intervals. Descriptive statistics may be contrasted with inferential statistics in that they simplify but do not attempt to extend beyond the immediate data (see [Huntsberger, 1961] and [Trochim, 2005]). Thus confidence bands are descriptive statistics used to summarize the developed probability densities; the bands do not extend the data provided by the densities. The density generation and characterization process is also applied to CEG curves, and it may be applied to other metrics.

The framework requires density models for target and non-target detection system output and prior densities for model parameters. The Bayesian approach incorporates all that is known or assumed about the data and density models. For a given set of target samples and non-target samples, assumed sample density models, and assumed prior densities, the Bayesian formalism permits the development of a ROC curve density. Other descriptive statistics, such as the ROC curve confidence intervals, may then be developed, if desired, from this probability density. Other methods focus up front on descriptive statistics (e.g., the mean and standard deviation of the target and non-target samples); such methods force premature simplification of the data; and either do not account for the model assumptions and priors densities or can only account for them in an ad hoc

manner. A Bayesian framework, by marginalizing over all possible models, provides a more robust estimate for a single set of data than other estimation methods. Methods other than Bayesian may perform well for large numbers of samples, but are less competitive for low numbers of samples. A ROC curve estimated by a maximum-likelihood approach is more accurate as the number of samples increases (see general discussion by [Robert, 2001, pp. 16]), but can not be relied upon for low numbers of samples. Non-Bayesian approaches can have superior performance if the Bayesian framework incorporates inappropriate model selection or prior density selection.

The Bayesian approach possesses two major strengths. First, it naturally and fully incorporates all possible model parameter values by marginalization (i.e., weighted averaging over all possibilities). The Bayesian approach avoids descriptive statistics until the parameters that are not of direct interest are integrated out and, thus, fully accounted for. In contrast, a maximum-likelihood approach attempts to find the "best" parameters (e.g., leading to a single best ROC curve). The maximum-likelihood based approach must then make additional assumptions (perhaps normal-based) to describe uncertainty. Bayesian approaches are more tolerant; the focus is not on finding a true single answer (see [Morgan, 1968, pp. 109]), but instead on describing the range of all possible answers in the form of a probability density, which is then more easily transitioned to other descriptive uncertainty statistics. Second, the Bayesian approach naturally incorporates the use of prior densities; that is, it permits the incorporation of subjective probability estimates into its framework, which is particularly critical when sample size is small (see [Good, 1965, pp. ix]).

Bayes estimators that perform point estimation, rather than the broader uncertainty estimation that is the focus of this research, are well known. Bayes' estimators can be fully consistent with traditional means of estimation, such as minimum mean square error (MMSE) and maximum a posterior (MAP) estimation (see [Scharf, 1991]). Robert [Robert, 2001] states that a Bayesian approach is consistent with three tests for

optimality from a non-Bayesian perspective: minimaxity, admissibility, and equivariance. Minimaxity typically consider the worst case scenario, but in contrast to frequentist-based approaches, a Bayesian approach prevents unwarranted reliance on a worst case scenario that has little chance of occurring (see [Robert, 2001, pp. 67], [Leonard and Hsu, 1999, pp. 146], [Schervish, 1995, pp. 167], and [Duda *et al*., 2001, pp. 28]). Admissibility focuses on whether or not there exists a better decision rule (see [Ferguson, 1967, pp. 54] and [Lehmann and Casella, 1998, pp. 323]) than the one selected. Equivariance relates to whether or not an estimate is invariant under linear transformation (see [Lehmann, 1998, pp. 161, 245]). Robert [Robert, 2001] shows that Bayesian estimators are a specific and preferred class of admissible estimators (see also [Schervish, 1995]).

For further discussion on the advantages of Bayesian-based approaches over more traditional methods, see [Good, 1965], [Schmitt, 1969], [Lindley, 1972], [Antelman, 1997], [Leonard and Hsu, 1999], [Robert, 2001], and [Woodworth, 2004].

2.6 Performance metric densities and confidence bounds

Figure 2.5 extends the relationships indicated in Figure 1.2 from simply identifying the performance metrics to formulation of probability densities of performance metric curves and values. It indicates three types of inputs: target and non-target samples, model specification, and sampling protocol.

As will be discussed in detail in Chapter 3, a reasonable model specification, if the sample scores are between zero and one, is a beta density. The beta density is specified by two parameters, mean and standard deviation. The model specification also includes prior assumptions for the parameters; for example, prior assumptions may be uniform prior densities for the mean and standard deviation over their allowed domains (the admissible set, defined in Chapter 3, specifies the allowed domains). Another example model is a truncated Gaussian density with uniform prior mean and variance (rather than uniform prior mean and standard deviation). The sampling protocol is also selected, but

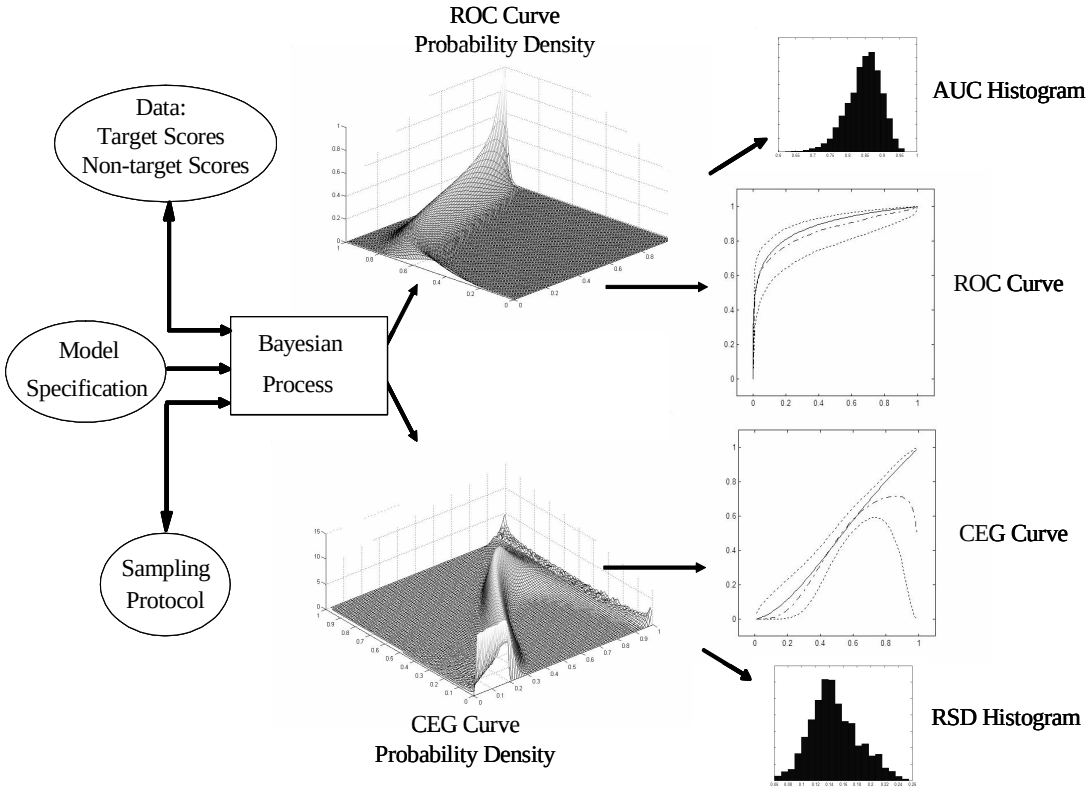


Figure 2.5 Uncertainty estimation process. Data (such as a set of 30 target and 30 non-target scores), Model Specification (such as beta probability densities for target and non-target and uniform prior densities for their means and standard deviations), and Sampling Protocol (such as uniform density of points from the prior densities), are inputs to a Bayesian Process. Outputs are probability densities for receiver operating characteristic (ROC) and confidence error generation (CEG) curves. These densities are characterized by plots that involve descriptive statistics, including histograms of area under receiver operating characteristic (AUC) and root square deviation (RSD) values from the ideal CEG curve, and also median ROC and CEG curves and corresponding curves that bound 90% of the probability density.

results which are not sensitive to this selection (and that approach an analytical solution; see Chapter 3) are obtained provided that a fine enough spacing in target and non-target parameter density is used. Monte Carlo methods may also be employed to generate points that sample the target and non-target density parameter values (see also Chapter 3).

The outputs of the Bayesian process (such a process accounts for all input data, prior densities, and integrates out free parameters through marginalization techniques) indicated in Figure 2.5 include performance metric densities (for example, ROC and CEG curve densities). The developed densities can be considered actual posterior probability densities (see [Carlin and Louis, 2000, pp. 35-36] for a discussion of actual probability statements) for the input samples (which are assumed independent and identically distributed for the research reported here), assumed density model, and prior densities of the model parameters. Although they are actual probability statements based on available samples, the developed probability densities are expected to change for more samples or different sets of samples. From a Bayesian standpoint, posterior probabilities are subjective and "quantify degrees of beliefs" (see [Mackay, 2003, pp. 26, 50]), so the developed posterior probability densities do not necessarily encompass truth if the model or priors are incorrect. Alternatively, if the selected model or priors are considered estimates, then the posterior probability densities may be considered estimates. Here, since the focus is on consistency with recent Bayesian literature, the term "posterior probability densities" rather than "posterior probability density estimates" is used. Note that Chapter 3 describes the performance metric density generation method, which is fully Bayesian in that it accounts for all assumptions and data and integrates out free parameters through marginalization. After performance metric densities are produced, probability density characterization produces descriptive statistics for the ROC curve, CEG curve, AUC value, and RSD value, as described and verified in Chapter 4. The four figures that form the rightmost column of Figure 2.5 show such statistics.

2.7 Literature review

Next, a literature review on methods for ROC curve estimation is presented; then a review on related confidence interval/confidence band methods is provided. Existing approaches have unacceptable weaknesses (e.g., they are only effective for large sample sizes, are restrictive to particular ROC curve shape, or make other unacceptable assumptions). The inadequacy of methods in the literature are identified here so that the benefits of the full Bayesian framework that are described in Chapters 3 and 4 can be better appreciated; the literature review provided here is not necessary to understand the method developed in Chapters 3 and 4. Later, Chapter 5 provides quantitative comparison of methods in the literature to the method that is developed here. Also, the CEG curve literature is reviewed; however, existing CEG curve literature does not provide adequate means of uncertainty estimation. The Metz [Metz *et al.*, 1998] method is examined as a primary example. Then, other methods of ROC curve estimation and ROC curve confidence interval estimation are examined.

2.7.1 Metz method. The Metz method, based on binormal ROC curve theory, is implemented in a software package called ROCKIT; ROCKIT is perhaps the most widely accepted ROC curve confidence interval software available today (see [Eng, 2005]). Binormal ROC curve theory assumes that the target and nontarget variables (referred to as diseased or non-diseased in the medical literature) are either normal or can be made normal after some unknown transformation. Binormal ROC curve development requires that, rather than plotting the ROC curve along correct detection probability and false alarm probability axes that are both uniform between zero and one, the axes use a linear scaling along normal deviate values, and this scaling is therefore non-uniform between zero and one [Dorfman and Alf Jr., 1968, 1969], [Swetz and Pickett, 1982], and [McNeil and Hanley, 1984]). Once the ROC curve is estimated as a straight line in normal deviate space, the ROC curve is then transformed into standard axes that are uniform between zero and one. Generally the curve, after being transformed into the

standard axes, has a convex appearance (although, as detailed later, the curve can have a "hook" that is especially apparent for small numbers of samples).

Historically, the binormal approach is the most common in the literature for rating scale data [Hanley, 1999]. Rating scale data are broken down into a number of distinct categories (typically five) in contrast to data described on a continuous scale. With five categories, five ROC points are plotted on the normal deviate plot described above. Upon conversion back to a scale that is uniform from 0 to 1 for both false alarm probability and correct detection probability, the line becomes the ROC curve. Note that because of assumptions due to plotting on the normal deviate axis, it is inappropriate to fit a least squares line to find the slope and intercept in the normal deviate space that best represents the ROC curve. Instead, a maximum likelihood method is used. Dorfman [Dorfman and Alf Jr., 1968, 1969] proposes a widely accepted method that estimates the ROC curve in such a manner. For an alternative maximum likelihood estimation development, see [Metz, 1984].

Metz [Metz *et al.*, 1998] develops an algorithm that extends the binormal approach to a large number of distinct categories, and therefore permits application of the binormal approach to a continuous scale.

Metz [Metz *et al.*, 1998] (and Swets [Swetz and Pickett, 1982]) alleviates the need to estimate the target and nontarget distributions directly. Metz found that the binormal approach provides satisfactory ROC fits to data generated in a "very broad variety of situations".

Here we consider what "broad variety of situations" means in a medical context. In the medical decision community, it is assumed that by measuring a known marker (from a blood test, for example) which indicates a disease, that the likelihood of disease in all cases is monotonically increasing (or decreasing) as marker level increases. For a target detection system under test, this is clearly not necessarily the case (while the monotonic

property is desirable for a system under test, one of the primary reasons for estimating the entire ROC curve is to determine if it is true, not to assume that it is true). Therefore, an assumed binormal ROC curve fit has weaknesses for target detection system evaluation.

Many applications that rely on binormal theory actually are interested primarily in the Area under the ROC curve (AUC) accuracy rather than the curve itself. The binormal ROC curve is a good estimate of AUC value, but is recognized as being of less utility when attempting to estimate an unknown ROC shape. Hajian-Tilaki [Hajian-Tilaki *et al.*, 1997] concludes that a binormal model is a robust method for determining AUC. However, they state that other indices, such as true-positive estimation fraction at a specific false-positive fraction point, might be more sensitive to departures from binormality.

The binormal ROC has recognized limitations, particularly for small numbers of samples. In general for many medical diagnostic scenarios, there is a large amount of sample data. So, requiring large sample sizes as a precondition may be reasonable for the medical community. The originator of binormal ROC maximum-likelihood theory, Dorfman [Dorfman *et al.*, 1997], states that the binormal ROC is not robust in small sample sets (Metz was a coauthor of the 1997 paper). Further, a study by Obuchowski [Obuchowski and Lieber, 1998] is unsupportive of the usefulness of a binormal ROC curve model (and other alternative ROC curve models) in estimating accurate confidence intervals in studies with small sample sizes.

Because of recognized inaccuracies in the binormal ROC when the true unknown ROC is assumed to be convex (the transformation from a linear plot in normal deviate space results in a ‘hook’ that can be particularly prevalent for small numbers of samples), Metz and Dorfman [Dorfman *et al.*, 1997] [Metz and Pan, 1999]advocate the development of a correction factor. Thus, it is recognized that even for the general assumptions for which binormal ROC theory is applicable, there are limitations. The desire to remove “the hook” has its origin in the assumption that the likelihood of observing a target increases

monotonically as the target score increases – i.e., the assumption that the appropriate model for a ROC is a convex shape. This assumption is not appropriate for ROC curves that evaluate a target detection system utility.

ROCKIT, which will be later used to provide in the course of comparisons with the method developed here, takes target and non-target sample inputs (either from user created files or from keyboard input). The user must specify whether such sample inputs be handled on a continuous scale or on a ratings scale, and the user must specify whether high or low scores values refer to targets. Then, ROCKIT produces an output file that contains estimates for points on the ROC curve (generally false alarm probabilities of 0.05, 0.01, 0.02, ..., 0.10, 0.20, 0.90, 0.95), AUC value, estimates for the binormal parameters that are used to form the ROC curve, 95% confidence intervals for the ROC curve, uncertainty estimates for the AUC value, and uncertainty estimates for the binormal parameters.

Chapter 5 provides a full comparison of the method developed here with the Metz approach described above. The weaknesses of the Metz method compared with the method developed here is even more apparent in the comparison provided by Chapter 5.

2.7.2 Other existing methods. Figure 2.6 diagrams methods in the literature which estimate ROC curves. The oval regions identify fundamental techniques that estimate ROC curves and compute ROC curve uncertainty, and the unshaded rectangular regions identify authors, years, and approaches. The shaded rectangular regions identify available ROC curve-related software, where the arrows to the software indicate the approaches they employ. Practical use of a SUT that is described by a ROC curve requires the selection of a threshold. Unless the underlying non-target density is deterministic, there is uncertainty in which false alarm probability corresponds with a particular threshold. Greenhouse [Greenhouse and Mantel, 1950] forms bounds to describe this type of uncertainty. Linnet [Linnet, 1987] extends this evaluation to ROC

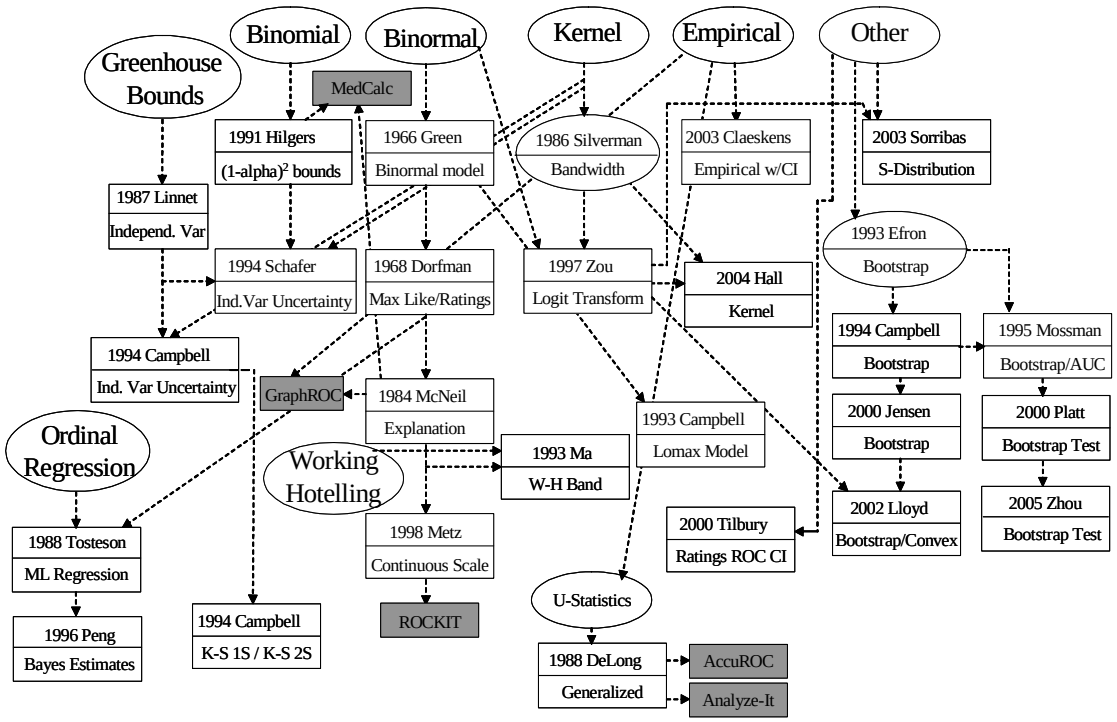


Figure 2.6 Relevant ROC curve literature and software. The figure shows an overview of relationships of ROC curve estimation and confidence interval development available in the literature. Underlying processes (not necessarily specific to ROC curves) are typically leveraged to estimate the form of ROC curves. The oval regions identify fundamental ROC curve estimation techniques (e.g., binomial, binormal, kernel, empirical). The estimation techniques permit the calculation of confidence intervals. The lines indicate relations among methods. The relations are only between the line origination points and the end points indicated by arrows. Several software packages, indicated by shaded boxes, apply particular ROC curve estimation processes and/or ROC curve uncertainty estimation processes (e.g., ROCKIT, MedCalc).

curves, and Schafer [Schafer, 1994] builds on work by Linnett and Wieand [Wieand *et al.*, 1989]. A disadvantage of the Greenhouse bounds is that such uncertainty in false alarm probability is assumed to follow a normal distribution.

Hilgers [Hilgers, 1991] details a method that generates confidence bounds for ROC curves based on binomial proportions. He applies ordered statistics to obtain confidence intervals given an interval range of interest (e.g., 90%) for each of a set of samples. For example, if there are five target samples, he estimates the lowest-valued sample for a two-sided 90% confidence interval to be between 0.02 and 0.53 of the overall cumulative distribution function (CDF) for target. He then estimates the second-lowest-valued sample for the same two-sided 90% confidence interval to be between 0.07 and 0.70 of the overall CDF for target. Finally, he combines the estimates to obtain confidence intervals for probability of correct detection and probability of false alarm. A constraint on the Hilgers approach is that the confidence intervals are "pointwise" and describe the range for a single point on the ROC curve. Hilgers extends these bounds to a confidence band by using a progression of rectangles based on the pointwise confidence intervals. However, Schafer [Schafer, 1994] shows that this procedure leads to an estimated bound larger than 90%. An advantage of the Hilgers approach is that it generates 'distribution-free' confidence bounds, unlike many approaches (most of which require some assumptions such as binormal target/non-target densities). Examples considered in Section 5.4 are consistent with Schafer in that the bounds are wide compared with the approach developed here.

Non-parametric approaches develop ROC curves analytically and do not assume a form for the underlying distributions. Zou [Zou *et al.*, 1997] provides an example which uses a Parzen window-like data transformation, referred to as kernel density estimation [Silverman, 1986]. Kernel density estimation enables ROC curve construction using a smoothed histogram. Zou leverages Silverman to describe the kernel density estimation of target or non-target density as

$$\hat{f}(x) = \frac{1}{mh} \sum_{i=1}^{m} k(\frac{x - X_i}{h}), \tag{2.13}$$

where k is the kernel density, m is the number of samples, X_i is the ith sample in $\mathbb{R}$, and $h > 0$ is the kernel width. Zou indicates that estimating f, in effect, places at each X_i in the sample an enclosed curve with area $1/m$, where each curve has a shape described by the function k and scaled by the width h. The curves are then added with the goal of obtaining a smooth but accurate histogram. With kernel density estimation, the function chosen for k is somewhat arbitrary, as is the selection of function width. Improved methods for width selection are desirable, but the optimization process is subjective. For example, Hall [Hall and Hyndman, 2003] explores methods for improving bandwidth selection, and Hall [Hall *et al.*, 2004] considers a method that makes width-dependent assumptions and generates results based on kernel estimation. The results of Hall show potential for significant degradation as false alarm probabilities approach 0 or 1 (these degradations are quantitatively compared with the method developed here in Chapter 5). Sorribas [Sorribas *et al.*, 2002] introduces a S-distribution that is related to kernel density estimation methods, and Campbell [Campbell and Ratnaparkhi, 1993] estimates ROC curves based on the Lomax distribution; neither approach introduces new methods of confidence interval development.

In principle, the goal of empirical approaches is to estimate ROC curves without making distribution assumptions. Claeskens [Claeskens *et al.*, 2003] is the most recent among many authors who consider empirical ROC curve estimation. As is typical, Claeskens recognizes the need for a smooth ROC curve and he uses kernel smoothing estimation, which thus introduces some distribution assumptions. Claeskens presents confidence regions for ROC curves with definitions similar to those of Hilgers that involve the regions of uncertainty for both correct detection probability and false alarm probability at a given threshold. Claeskens discusses other confidence interval descriptions, but reverts to a bootstrap confidence interval estimation method when these confidence intervals are

calculated. Earlier approaches in the empirical category are considered by Hsieh [Hsieh and Turnbull, 1996], and a local smoothing technique is investigated by Qiu [Qiu and Le, 2001]; however, neither approach fully develops confidence intervals.

Ma [Ma and Hall, 1993] applies the Working-Hotelling hyperbolic confidence band for multiple regression surfaces to ROC curves; they generate pointwise confidence bands by varying correct detection probability and mapping a band of intervals for false alarm probability and also simultaneous confidence bands for the entire ROC curve. Some limitations of this approach are that the confidence bands for the entire ROC curve assume binormality, and the method uses rating scale data. However, their approach extends to multiple confidence interval and confidence band definitions, and they emphasize the need for such definition flexibility. Although Ma claims that the Working-Hotelling approach extends beyond binormal methods, confidence bands are obtained using conventional binormal assumptions applied to ratings scale data. Further, the Working Hotelling approach applies only when the assumptions made permit the use of regression lines.

Confidence intervals may be generated using various resampling methods, even if different methods develop the ROC curve estimates. Examples are in [Zhou and Qin, 2005], [Platt *et al.*, 2000], [Jensen *et al.*, 2000], [Mossman, 1995], [Campbell, 1994], [Garber *et al.*, 1994], and [Simpson *et al.*, 1989]. Efron [Efron and Tibshirani, 1993] details general bootstrap theory that is often leveraged in ROC curve resampling processes (see Mossman [Mossman, 1995] and Jensen [Jensen *et al.*, 2000]). The confidence interval results are generally jagged in appearance (as shown in Figure 5.3 of Chapter 5), and the coverage areas are inaccurate for low numbers of samples, particularly in regions of low correct detection probability density.

Lloyd [Lloyd, 2002] implements bootstrap confidence methods by evaluating ROC curve definitions many times in a Monte Carlo approach. He obtains confidence intervals using a maximum likelihood approach to estimate the ROC curves parametrically and

non-parametrically. He does not verify the coverage accuracy of the bootstrap method, and he cautions that bias may be a significant disadvantage for small samples.

Once target and non-target data are obtained, Tilbury [Tilbury *et al.*, 2000] asks for every point on the ROC curve, "If this point represents the true Hit Rate and False Alarm Rate of the population, what would be the probability of getting the sample actually obtained." He analyzes one point (false alarm probability and correct detection probability at one selected threshold) on the ROC curve, then he considers a combined approach for four selected thresholds. For just four points, he obtains a solution based on an eight-dimensional hyperboundary, where increasing the number of initial points on the ROC curve increases the dimensions needed. He suggests estimating ROC curve density by selecting a point on the ROC curve and finding the likelihood that given samples (assuming a threshold) are generated if this point is from the underlying densities (consistent with Hilgers-like binomial based approach). Tilbury requires an expansion of dimensionality based on the number of samples.

Although Tilbury's approach remains tractable if a few selected thresholds are permitted (through grouping of data), Macskassy [Macskassy and Provost, 2004] declares Tilbury's method not tractable for more than ten points. Tilbury provides updates to his work [Tilbury 2002, 2003a, 2003b] that emphasize the importance of Bayesian statistics in ROC curve analysis, and he uses Bayes' rule in considering the descriptions of the 2000 paper. However, his approach remains a binomial-based alternative to Hilgers' [Hilgers, 1991] approach. Tilbury [Tilbury *et al.*, 2000] claims verification of results for uncertainty of correct detection probability and false alarm probability, but these are (at best) simply verified coverages for single thresholds considered independently (even here, he does not report actual accuracies, but provides tables of distributed data, and he does not compare results with other research). Tilbury's method in theory permits incorporation of prior densities of false alarm and correct detection probability, but not prior target and non-target densities.

In summary, Tilbury provides an alternate description of the work of Hilgers [Hilgers, 1991] by leveraging binomial assumptions and forming contour regions for particular thresholds rather than the rectangular regions of Hilgers. His method does not permit the incorporation of different target and non-target density models or target and non-target prior parameter densities, and he does not demonstrate the practical development of a ROC curve confidence band jointly across the entire curve (such as those that are tested for coverage accuracy in the research reported here). His method produces confidence bands for particular thresholds similar to Hilgers but with different shape. Tilbury attempts analytically to show how such regions could be combined, but he avoids verification (consistent with Macskassy's tractability concerns), except for correct detection probability and false alarm probability uncertainty regions at individual threshold points (similar to Hilgers). Further, his approach is based on proportions that correspond with the correct detection and false alarm probability models but do not correspond directly with "score" and "probability of target given score". Thus, Tilbury's ROC curve confidence interval approach does not extend to the CEG curve and other performance metrics.

Tosteson [Tosteson and Begg, 1988] develops regression parameters to estimate the shape of the ROC curve for a fixed number of thresholds (such as five thresholds). The regression parameters attempt to describe the relation of covariates such as stage of disease, age, and weight to the estimated ROC curve. Several related extensions develop Bayesian-based approaches to more robustly account for the regression parameters (see [Peng and Hall, 1996], [Hellmich *et al.*, 1998], and [Zou and O'Malley, 2005]). These approaches assume a binormal ROC curve form. Smith [Smith *et al.*, 1996] provides an alternative to the binormal-based methods, but Smith's approach also makes curve shape assumptions. O'Malley [O'Malley *et al.*, 2001] provides an alternative to the grouped data methods but still makes binormal assumptions. Each of these regression based approaches have significant limitations compared with the method developed here. The methods are restricted to an assumed shape of the curve; a shape is not assumed for the

SUT-focused ROC curve estimates developed here. Due to the focus on shape parameters, the methods are not generally transferable to other performance metrics such as the CEG curve. Further, the efforts do not consider confidence interval coverage accuracy verification. Zou [Zou and O'Malley, 2005], O'Malley [O'Malley *et al.*, 2001], and Smith [Smith *et al.*, 1996] avoid ROC curve confidence intervals altogether, and Hellmich [Hellmich *et al.*, 1998] and Peng [Peng and Hall, 1996] provide confidence intervals based on the binormal mean and slope parameters but do not verify coverage accuracy. Note that the methods listed above focus on alternatives to maximum likelihood estimation for generally binormal based ROC curves rather than uncertainty in such estimates.

A number of authors leverage Bayesian approaches in order to combine ROC curve results for meta-analysis applications; meta-analysis focuses on pooling the results of multiple diagnostic tests (see [Carlin, 1992], [Smith *et al.*, 1995], [Zhou, 1996], [Hellmich *et al.*, 1999], [Rutter and Gatsonis, 2001], and [Dukic and Gatsonis, 2003]). Such approaches use Bayesian-based processes to combine the ROC curves and AUC value of each individual test into a combined estimate of the underlying true ROC curve and AUC value.

Various approaches focus solely on AUC value uncertainty (see [DeLong *et al.*, 1988], [Broemeling, 2004], [Yousef *et al.*, 2005], [Agarwal *et al.*, 2005], and [Cortes and Mohri, 2005]). DeLong [DeLong *et al.*, 1988] leverages U-Statistics to provide an estimate of whether two AUC values are statistically different from one another; DeLong includes an evaluation of uncertainty in making such estimates. Yousef focuses on AUC value standard deviation (which may exceed one) as a description of uncertainty. Yousef's approach has limitations, as the AUC values may be skewed and must be less than one. Yousef does not have a true verification process, only a comparison with results that are already available through traditional bootstrapping processes. Yousef assumes that ROC curves have convex form. Agarwal and Cortes

develop approaches that focus on uncertainty in the Mann-Whitney statistic (the Mann-Whitney statistic enables computation of the AUC value without development of an entire ROC curve), and both methods are limited to large numbers of samples. In comparison, the method developed here focuses on ROC curve uncertainty, although the results are also successfully applied to AUC value uncertainty and then extended to CEG curve uncertainty. Broemeling proposes a Bayesian based approach to AUC value estimation, but his method is only applicable for a limited, fixed number of possible thresholds (Broemeling uses five thresholds), rather than the continuous set of possible thresholds that the research developed here makes possible. Broemeling computes AUC value confidence intervals for two examples but does not verify coverage accuracy.

Dass [Dass and Jain, 2005] provides an approach to ROC confidence bands but with a focus on correlated samples. The Dass approach is restricted to correlated samples (rather than independent samples), is limited to large numbers of samples, and does not verify coverage accuracy.

Overviews of ROC curve theory are given by Centor [Centor, 1991], Hanley [Hanley, 1999], and Zweig [Zweig and Campbell, 1993]. Hanley and Zweig provide relevant overviews in the ROC curve confidence interval area. More recently, Macskassy [Macskassy *et al.*, 2005][Macskassy and Provost, 2004] reviews ROC curve confidence interval approaches for the machine learning community, and Carsten [Carsten *et al.*, 2003] evaluates ROC-curve-related software. Bamber [Bamber, 1975], Lusted [Lusted, 1971], and Swets [Swetz and Pickett, 1982] provide historical background on ROC curve theory. Bamber identifies the underlying purpose and meaning of AUC value. Lusted summarizes the origins of ROC curve theory as related to signal detectability. Swets and Pickett provide a widely recognized reference text on ROC curve theory. Green [Green and Swets, 1988] provides a detailed ROC theory review in a reprint/revision of a text originally written in 1966. Lusted discusses the

relation of the medical decision making and radar progression in ROC curve development [Lusted, 1984].

As mentioned in Section 2.2, in medical research sensitivity is typically used in place of correct detection probability, and one minus specificity replaces false alarm probability. Similarly, "diseased patients" often replaces "target data", and "healthy patients" replaces "non-target data". The discussion here refers to target, non-target, probability of correct detection, and probability of false alarm for consistency even when the literature uses different (but analogous) terms.

Figures 2.7, 2.8, and 2.9 provide an overview of existing ROC curve confidence interval approaches. A review of the approaches listed in these figures reveals differences in confidence interval definitions and emphasizes that existing methods lack robustness and flexibility, the methods typically identified in the research are focused on a subset of the possible confidence bound definitions and do not extend to other definitions. Confidence bound definitions are summarized as follows.

Confidence definition 1: fixed threshold. This definition selects a particular threshold, develops an estimate for false alarm probability uncertainty, and similarly develops correct detection probability uncertainty. Approaches in the literature often attempt to extend this approach. For example, a rectangular region is created based on the uncertainties in false alarm and correct detection probability. A complete estimate of ROC curve uncertainty is then made by connecting the corners of the boxes (see Figure 5.9). A weakness of this ad hoc approach is that typically the confidence interval band is wide compared with other approaches, particularly at low sample sizes. Examples are considered by Hilgers [Hilgers, 1991].

Confidence definition 2: uncertainty in correct detection probability at given false alarm probability. This definition regards false alarm probability as the independent variable,

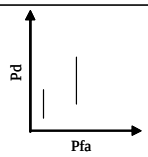
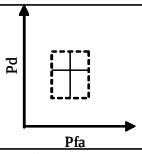
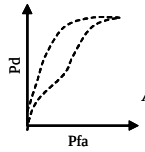
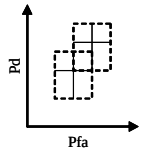
		Confidence Definition	Distribution Assumptions	Confidence Example or Verification	Comments
1987 Linnet Indepen. Var		*Definition 2:* Uncertainty in Pd at Pfa	Normal and non-parametric (but symmetric) methods	Normal example, No verification	First ROC recognition of Pfa uncertainty
1991 Hilgers $(1\text{-alpha})^2$ bounds		*Definition 1:* Fixed Threshold *Also: Definition 4:* Full Curve	Binomial Order statistics	Example, No verification on full curve, Individual bounds verified [Ross, 2003]	Large confidence band area
1993 Ma & Hall W-H Bands		*Definition 4:* Full Curve *Also: Definition 1, 2.*	Binormal, Working-Hotelling Regression theory	Binormal example, No verification,	Emphasize use of multiple confidence definitions
1994 Campbell Ind. Var Uncertainty		*Definition 2:* Uncertainty in Pd at Pfa	Kolmogorov distribution theory	Example, No verification	Confidence bounds made up of same size rectangles

Figure 2.7 ROC literature comparison I. Confidence interval approaches are listed by author. Correct detection probability is Pd and false alarm probability is Pfa. The first column lists confidence interval or band definitions. The second column lists distribution assumptions. The third column indicates whether confidence interval examples or verified results are provided. The most promising verified results are compared with the method developed here in a later section. The fourth column comments on significant attributes of the corresponding research.

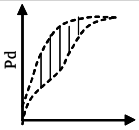
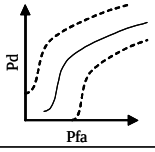
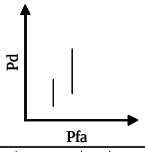
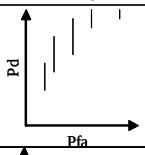
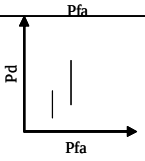
		Confidence Definition	Distribution Assumptions	Confidence Example or Verification	Comments
1994 Schafer Ind.Var Uncertainty		*Definition 2:* Uncertainty in Pd at Pfa	Asymptotic theory	Normal, Binormal examples, verification	Coverages large. unreliable for small samples
1994 Campbell Bootstrap		*Definition 4:* Full Curve *Also: Definition 1, 2.*	None (bootstrap resampling)	Example, No verification	Symmetric Bands, Fixed width displacement
1997 Zou Logit Transform		*Definition 2:* Uncertainty in Pd at Pfa	Kernel logit transformation	Beta mixture model example, No verification	
1998 Metz Continuous Scale		*Definition 2:* Uncertainty in Pd at Pfa	Binormal	Example, No verification	Implemented in 'Rockit' software
2000 Platt Bootstrap Test		*Definition 2:* Uncertainty in Pd at Pfa	None (bootstrap resampling)	Beta and normal verification	Verifies Linnet and other approaches

Figure 2.8 ROC literature comparison II. For explanation, see Figure 2.7.

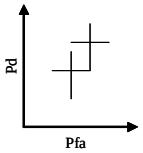
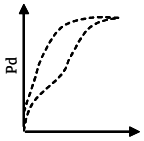
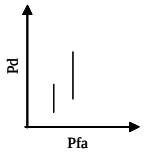
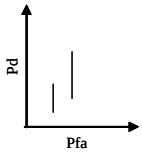
	Confidence Definition	Distribution Assumptions	Confidence Example or Verification	Comments
2003 Claeskens Empirical w/CI 	*Definition 1:* Fixed Threshold	Kernel	Verification only specific thresholds	
2003 Claeskens Empirical w/CI 	*Definition 4:* Full Curve	Empirical log-likelihood ratio to estimate curve, boostrap method for confidence band	No verification	Bootstrap bands based on Lloyd (1998)
2004 Hall Kernel 	*Definition 2:* Uncertainty in Pd at Pfa	Kernel	Verification	No confidence interval widths
2005 Zhou Bootstrap 	*Definition 2:* Uncertainty in Pd at Pfa	Adjusted binomial adjustment to bootstrap	Verification	Confidence interval widths

Figure 2.9 ROC literature comparison III. For explanation, see Figure 2.7.

and the literature and the method developed here tend to focus on this choice. Thus uncertainty in correct detection probability is calculated at a given false alarm probability, and confidence contours covering the entire ROC curve are developed by repeating for all given false alarm probabilities. Linnet [Linnet, 1987] notes that assuming false alarm probability is known when it is, in fact, uncertain introduces error in correct detection probability. Examples are considered by Campbell [Campbell, 1994], Linnet [Linnet, 1987], Schafer [Schafer, 1994], Zou [Zou *et al.*, 1997], Metz [Metz *et al.*, 1998], Platt [Platt *et al.*, 2000], and Zhou [Zhou and Qin, 2005].

Confidence definition 3: uncertainty in false alarm probability at a given correct detection probability. This approach is similar to confidence Definition 2, except that correct detection probability is regarded as the independent variable. For beta target and non-target densities, the method developed here produces confidence bands by this definition that are similar to the bands of confidence Definition 2. There are no known methods in the literature that focus on this method.

Confidence definition 4: full curve confidence band. This band represents the uncertainty of the entire ROC curve. The literature focuses less on this definition than on that of Definition 2. Examples are considered by Ma [Ma and Hall, 1993], Claeskens [Claeskens *et al.*, 2003], and Campbell [Campbell, 1994]. Bands by this method typically have the objective of enclosing the entire true ROC curve with a selected percentage confidence. If even a small portion of the ROC curve is outside of the band, then the entire band is regarded as being in error.

Confidence definition 5: curve location based on uniform threshold. This confidence bound describes ROC curves for a threshold chosen uniformly at random. Such bounds are not described in the literature but are a natural extension of the method developed here. Figure 4.10 shows ROC curve confidence bounds based on this definition and

shows higher densities close to the ROC curve extremes. This result is appropriate because any ROC curve has a correct detection probability of zero at a false alarm probability of zero and similarly a correct detection probability of one at false alarm probability of one.

Unlike the references to ROC curves, many CEG curve and RSD value approaches differ from those developed here. Since the metrics differ, the methods of obtaining confidence intervals or variance for the metrics also differ. For example, Lombard [Lombard, 2003] details an approach for estimating uncertainty in on-line gauges, O'Connor [O'Connor *et al.*, 2001] describes the asymmetry of confidence intervals related to weather forecasting, and Yaniv [Yaniv and Foster, 1997] analyze the precision and accuracy of judgmental estimation. The performance metrics described in the latter can be transitioned to confidence-error-like performance metrics.

The scores from a SUT are posterior probability estimates as detailed by Bishop [Bishop, 1995]. However, for the CEG curve the intent is not to estimate posterior probability but rather to estimate how well an unknown "black box" performs in providing estimates of posterior probability. Thus, the intent is to provide confidence intervals for CEG curve and RSD values, which characterize score posterior probability. El-Jaroudi [El-Jaroudi, 1990], Lugosi [Lugosi and Pawlak, 1994], Poggio [Poggio *et al.*, 2004], and Tomasi [Tomasi, 2004] focus on estimating error in posterior probability. Existing research is more relevant in formulating alternative approaches for determining confidence error than in quantifying confidence intervals, variance, and/or the density of confidence error. Also, another confidence-interval-like method involves cross-entropy (see [Bishop, 1995]), which is a metric often used in speech processing.

Research in the ATR community for performance metrics and confidence error includes work by Ceritoglu [Ceritoglu *et al.*, 2003], DeVore [DeVore, 2004], Irvine [Irvine *et al.*, 2002], Li [Li *et al.*, 2001], Mossing [Mossing and Ross, 1998], [Ross et al., 1997, 1998, 1999, 2002], , [Ross and Mossing, 1999], and Thorsen

[Thorsen and Oxley, 2004]. These references help identify the relevance and need for confidence error as a performance metric. Ross [Ross and Minardi, 2004] develops the rationale for a CEG curve-based performance metric and identifies the ability of such a metric to provide information on the performance of a target recognition system that the ROC curve is not able to provide. Ross points out that confidence errors (to include additional confidence measures of performance) are in themselves estimates and emphasizes that the ATR community needs confidence intervals for these estimates.

The underlying methods and techniques and probability density estimation methods that are leveraged to form ROC curve confidence intervals and CEG curve confidence intervals in the chapters that follow must be considered. The methods developed here apply a Bayesian framework to ROC curve and CEG curve performance metrics. A similar framework was devised in the early 1990s for neural networks applications [MacKay, 1992a, 1992b]; this framework has not heretofore been comprehensively applied to target detection performance metrics. Bishop [Bishop, 1995] provides a summary of MacKay's contributions. A critical aspect of the Bayesian approach is correct modeling of the prior parameter densities. For the beta density model considered here, it is shown that sampling uniformly over the domain of all means and standard deviations yields appropriate results. Chapter 3 describes the analytical convergence of this procedure, which may also be obtained using a Monte Carlo approach. As model parameters become more complex, other Monte Carlo methods and Bayesian techniques may be suitable alternatives to sampling uniformly over parameter domains. Clyde [Clyde, 1999] identifies search methods for posterior densities; and Clyde [Clyde and George, 2004] details advancements that make such posterior density searches practical. Barbieri [Barbieri and Berger, 2004] suggests a robust posterior density approximation that considers only parameter values which have posterior density weights that are 50% of the maximum posterior weight. Jordan [Jordan *et al.*, 1999] details various computational methods for calculating posterior densities. Hoeting [Hoeting *et al.*, 1999], Raftery [Raftery *et al.*, 2003], and Madigan

[Madigan and Raftery, 1994] discuss the application of Occam's razor, which refers to the concept that whereas more complex models are possible, the posterior density contribution of a simpler model should generally outweigh a more complex model (other constraints being equal). Occam's window reflects the concept that all parameter values that have less than a selected percentage of the maximum weighting can be disregarded without loss of accuracy [Hoeting *et al.*, 1999].

For the research presented here, the beta density is appropriate because this density is non-zero for score values between zero and one, and a single beta density has a simple unimodal form. However, the use of the beta density is also justified because it is the density of maximum entropy which is zero beyond a limited domain subject to two constraints, which may be related to the density mean and variance. Gokhale [Gokhale, 1975] investigates the usefulness of maximum entropy distributions subject to various constraints, and Kagan [Kagan *et al.*, 1973] documents the properties of the beta density relative to maximum entropy. Note also that several recent ROC confidence interval papers (see [Platt *et al.*, 2000], [Hall *et al.*, 2004], and [Zhou and Qin, 2005]) use beta densities to generate samples.

2.7.3 Summary of existing research. Each of the ROC curve uncertainty estimation methods discussed above have weaknesses that the method developed here largely overcomes. Some methods [Zhou and Qin, 2005] only provide acceptable results as sample size becomes large, which is the opposite of what is needed for the target detection applications considered here. Others methods are restricted to normal-based assumptions and can not be extended to other density forms (see [Ma and Hall, 1993]); binormal based approaches [Metz *et al.*, 1998] make unacceptable restrictions on functional forms. Still other methods (e.g. [Hilgers, 1991]) produce confidence regions that are too large and therefore uninformative. Further, most of the authors identified here refrain from quantitative verification of results; the few that do are examined in detail in Chapter 5. The quantitative comparison provided in Chapter 5 of the method

developed here with existing methods reinforces the above discussion. Recent literature in the ATR community introduces the basis for the CEG curve and RSD value described here, however, methods for their confidence interval (or band) uncertainty estimation are not available, although the need for such methods has been identified (see [Ross and Minardi, 2004]).

Thus, a review of the previous research reveals that a new method for performance metric uncertainty estimation is needed. The method developed and verified in Chapters 3 and 4 introduces a flexible new framework that can be applied to ROC curves and CEG curves, and it provides uncertainty estimates for these curves (and for their summary metrics of AUC value and CEG value).

3. Probability Density Generation

This chapter develops methods that generate probability densities for target detection performance metrics, such as the ROC curve. The development process has the following rationale. First, consider that deterministic performance metrics (e.g., a fully specified ROC curve with no uncertainty) assume that the target and non-target sample densities of score are known. Such exact target and non-target sample densities could be determined from the samples if it were possible to generate an infinite set of target and non-target samples. From a finite set of samples, it is not possible to determine exactly the target sample density and the non-target sample density. Thus, a set of possible densities for a finite set of samples is examined, with each density defined by values of one or more parameters (for example, the parameters for a univariate Gaussian density consist of mean and variance). Next, using a Bayesian process, parameter values for the target and non-target densities are found. Finally, the resulting densities of target and non-target samples are used to find probability densities for the performance metrics. The procedure for developing densities is applicable to any parametric density model (the beta density model is the example emphasized here). Once the performance metric probability density is generated, a variety of standard descriptive statistics may be developed, including mean, median, mode, confidence bounds, etc. Chapter 4, Probability Density Characterization and Verification, considers these descriptive statistics.

3.1 Target and non-target samples, density models, and ROC curve estimates

Section 2.2 focused on deterministic ROC curves, where the underlying target and non-target densities are known. This section focuses on the relation of samples to assumed underlying target and non-target score probability densities and on ROC curve estimates obtained from these densities. Figure 3.1 shows example target and non-target

densities, a set of samples generated from the target density, and a second set of samples generated from the non-target density. Here 30 target score samples (triangles) and 30 non-target score samples (circles) are drawn from their respective specified underlying densities. For an infinite set of such samples the target and non-target densities are known. In this ideal case the associated performance metrics of ROC curve, AUC value, CEG curve, and RSD value are deterministic and have no uncertainty. When only a finite number of samples are available, the target and non-target densities for an infinite number of samples are not known but are desired. Any density that is non-zero at each of the sample values has some probability of being the density formed by an infinite set of samples. However, it is appropriate to consider only density functional forms or models that incorporate additional available information, such as that density is continuous and is non-zero only between zero and one.

Beta densities are used to implement the performance metric uncertainty estimation framework developed here. While this density model is reasonable, a major advantage of the framework developed here is that it is applicable to other models. The beta density is of interest because it has zero magnitude outside the interval [0,1], as assumed for the target and non-target score data. Additionally, the beta density (see [Patel *et al.*, 1976] and [Mendenhall *et al.*, 1990]) has maximum entropy among all continuous densities that are non-zero only between zero and one and that meet two additional constraints [Kagan *et al.*, 1973] which may be related to mean and variance. The beta density with parameters $a, b > 0$ is

$$f(s) = \begin{cases} C_{a,b} s^{a-1}(1-s)^{b-1}, 0 \leq s \leq 1, \\ 0, \text{ elsewhere,} \end{cases} \tag{3.1}$$

where s is score and the mean and variance of the beta density are related to a and b by

$$\mu = a/(a+b) \text{ and } \sigma^2 = (ab)/[(a+b)^2(a+b+1)], \tag{3.2}$$

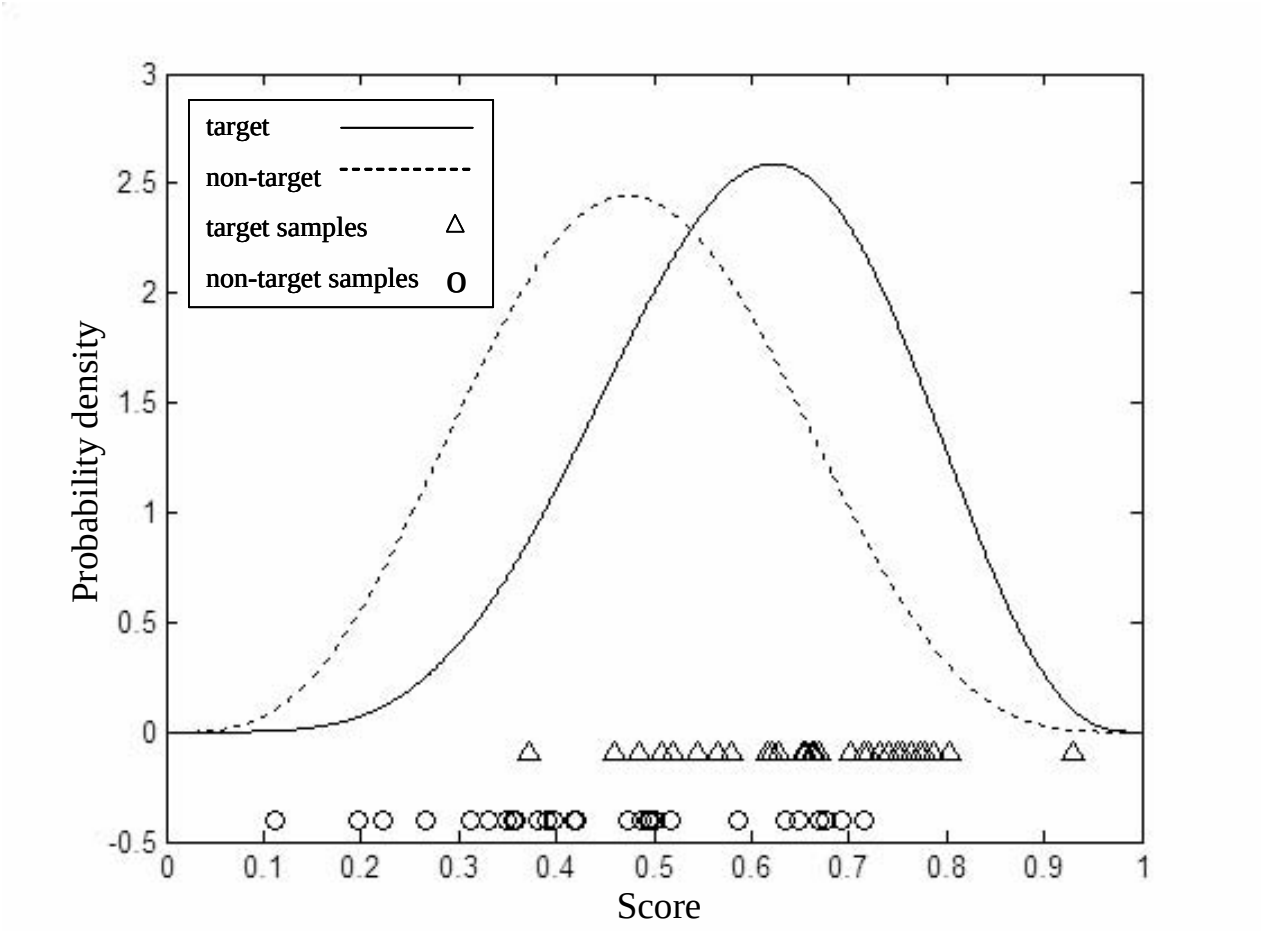


Figure 3.1 Target and non-target samples and the densities from which they are drawn. A target beta density (solid line) and a non-target beta density (dashed line) are shown; these densities are typically estimated from samples. Here 30 target score samples (triangles) and 30 non-target score samples (circles) are drawn from their respective densities.

and the constant $C_{a,b}$ equals $1/\int_0^1 s^{a-1}(1-s)^{b-1}ds \equiv 1/\text{Beta}(a,b)$.

A simple method for mapping a set of sample scores to a beta density is to find the mean and variance of the scores, then use them in Equation (3.2) to obtain the a and b values. Once the scores are mapped to beta density form (one density for target samples and the other density for non-target samples), a ROC curve and corresponding AUC value, as well as a CEG curve and a corresponding RSD value, are calculated. Note that using sample mean and variance to estimate a beta density, where the sample variance is unbiased in that it is the sum of squared deviations from the mean divided by the number of samples minus one, is equivalent to a maximum-likelihood approach as sample size increases (see [Hahn and Shapiro, 1967]).

Figure 3.2 compares ROC curve estimates for 10 sets of 30, 300, 1000, and 3000 target and non-target samples. To obtain such sets for comparison with the true ROC curve, first choose an underlying target density and non-target density. Then find the ROC curve that corresponds with these densities from Equation (2.10). This ROC curve, computed numerically, is shown as the solid line on each of the four plots. From the densities, randomly and independently draw 30 target samples and 30 non-target samples to obtain one set of data. Estimate the target and non-target beta densities as the densities with the mean and unbiased variance of the target and non-target samples (mean and variance determine the density parameter vectors u and v of Equation (2.10)), and form a ROC curve from these estimates. Find the 10 sets of ROC curves for the 30 target and 30 non-target samples, then repeat for 10 sets of 300, 1000, and 3000 pairs of target and non-target samples. Note that even for the 3000 sample example, differences in the ROC curve estimates are apparent. Figure 3.3 shows a similar progression, except that here the ROC curves are formed by evaluating the correct detection probability and false alarm probability at every score value using only the sample values and not an assumed model. Figures 3.2 and 3.3 indicate that ROC curve estimates for low numbers of samples may not be close to the true ROC curve. The variance shown in the plots in

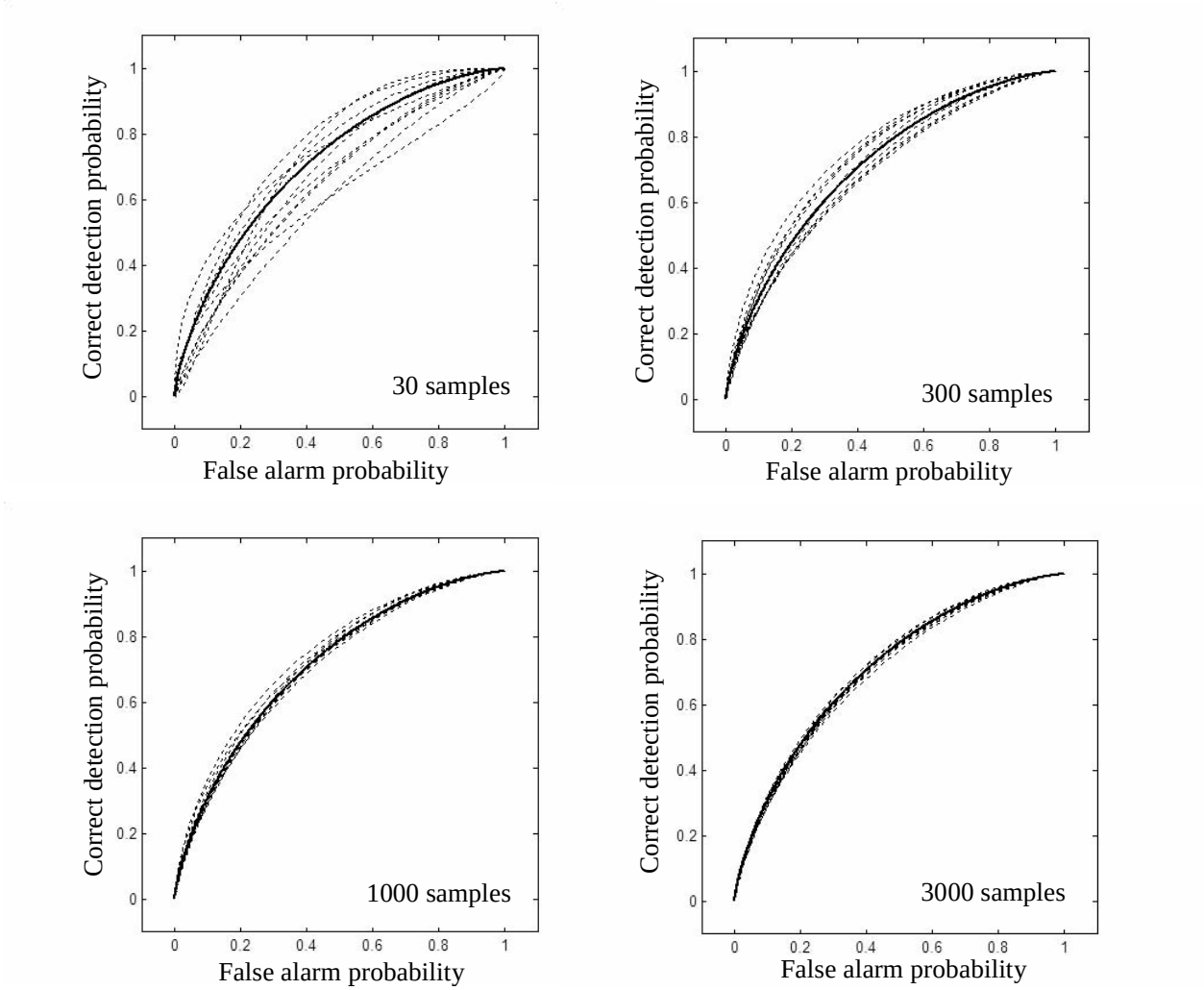


Figure 3.2 The ROC curve estimates for various sample sizes, where beta density estimates generate the ROC curves. Target and non-target beta densities generate target and non-target samples, and ROC curve estimates are formed from beta densities that have the mean and variance of the samples. For the top left plot, 10 ROC curves (dashed lines) for 10 sets of 30 target and 30 non-target samples are generated by fitting beta densities to the samples. In the other plots, similar sets of ROC curves for 300, 1000, and 3000 pairs of target and non-target samples are generated. The actual ROC curve that the densities form for an infinite number of samples is shown as the solid line on each plot. Variance is apparent in the plots, even for 3000 target samples and 3000 non-target samples.

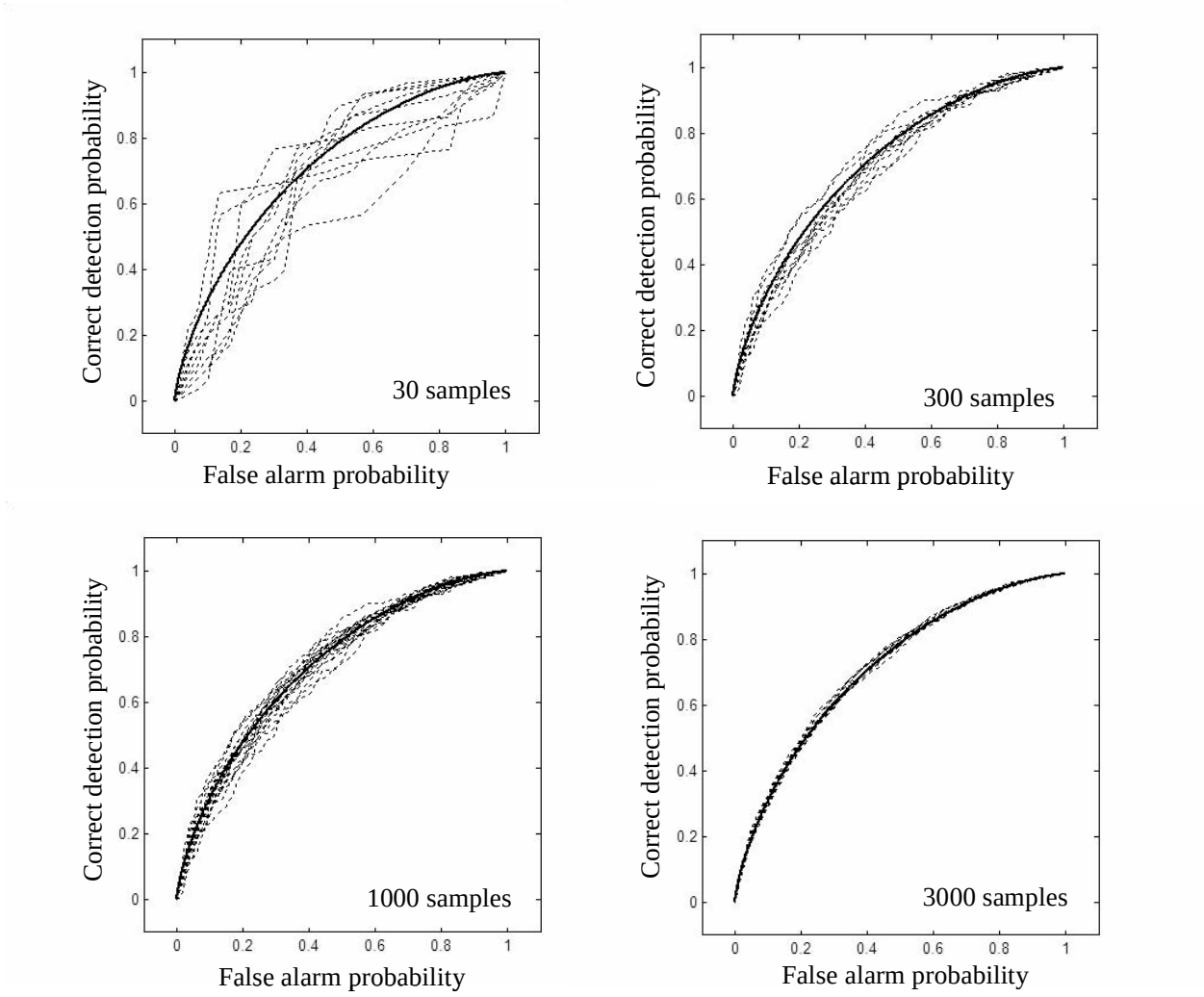


Figure 3.3 The ROC curve estimates for various sample sizes, where the empirical samples generate the ROC curves. The four plots are formed using the process of Figure 3.2, except the ROC curves are formed directly using the sample values; a beta density form is not assumed. The variance in each of the plots emphasizes the importance of ROC curve uncertainty estimation and the inadvisability of focusing on one ROC curve estimate.

these two figures emphasize the importance of ROC curve uncertainty rather than the estimated ROC curve. Section 3.2 details a fully Bayesian process for estimating ROC curve uncertainty.

Unimodal beta densities and score-threshold ROC curves are the assumed model and performance metric for much of the research discussed here; however, the beta density is used for illustration. The framework developed in the next section (with beta densities) may be applied to other density models and to likelihood threshold ROC curves. For example, multi-modal beta mixture models and related empirical-threshold and likelihood-threshold ROC curves are shown in Figures 3.4 and 3.5.

3.2 Bayesian posterior densities of parameters and weighted ROC curves

The two left plots of Figure 3.6 show the collection of pairs of means and standard deviations for beta densities that are zero at scores of 0 and 1. Values of standard deviation outside each “rounded triangle” do not exist for these densities. Values of standard deviation inside each "rounded triangle" are the admissible set, where the admissible set is described as follows. For the case of this beta density model, the admissible set $\mathcal{A}$ consists of (μ, σ) pairs such that

$$\left\{ \begin{array}{c} \text{if } 0 \leq \mu \leq 0.5,\ \sigma \leq \frac{1-\mu}{\mu(\mu+2)(\mu+1)^2} \\ \text{if } 0.5 \leq \mu \leq 1,\ \sigma \leq \frac{\mu(1-\mu)^2}{2-\mu} \end{array} \right\}. \tag{3.3}$$

Admissible sets may also be defined for other density models, including density models that are not restricted to two parameters. The target and non-target densities shown in the right plot of this figure map to unique locations on the standard deviation versus mean graphs shown at the left. Applying Bayes’ rule in a process consistent with that developed by [MacKay 1992a, 1992b] for the neural network community, but not heretofore applied to target detection performance metrics, the densities of model

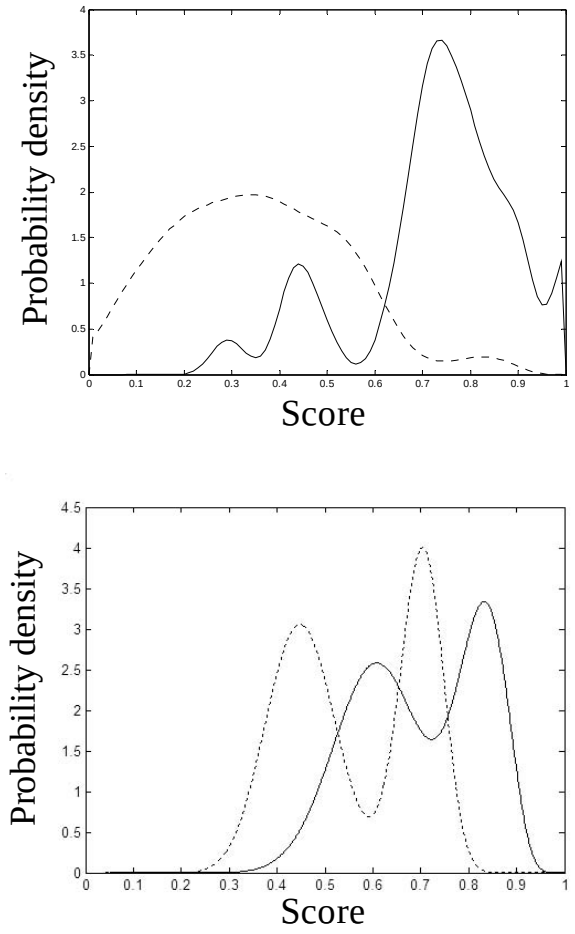


Figure 3.4 Target (solid) and non-target density (dashed) examples with a beta mixture model. In the upper graph two separate sums of 30 beta densities form the target and non-target densities. Similarly, a sum of two beta densities form each density in the lower graph. (The target density has 0.82, 0.055 and 0.7, 0.045 for the mean and standard deviation of the two beta densities, and the ratio of their amplitudes is 0.45. The corresponding five values for the non-target density are 0.6, 0.084, 0.45, 0.071, and 0.45).

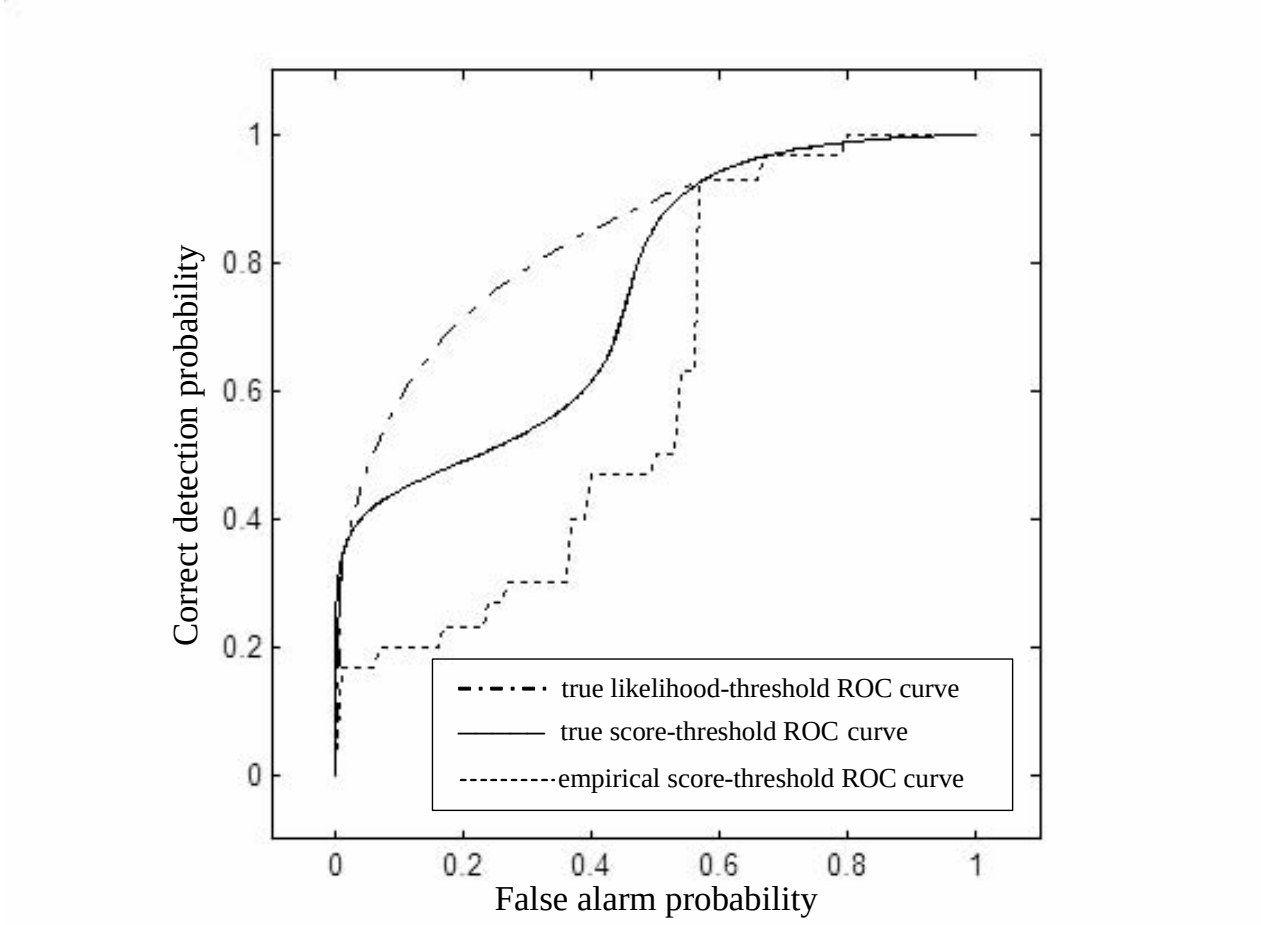


Figure 3.5 Relation of the true likelihood-threshold ROC curve (dot-dash line), the true score-threshold ROC curve (solid line), and the empirical-threshold ROC curve (dashed line). The true ROC curves assume knowledge of the underlying densities (shown in the the bottom plot of Figure 3.4). For the true likelihood-threshold ROC curve, probability of detection is the integral of the target density over the region to the left of the first vertical line and the region to the right of the the second vertical line in Figure 2.1. Similarly, the probability of false alarm is the integral of the non-target density over the same regions.

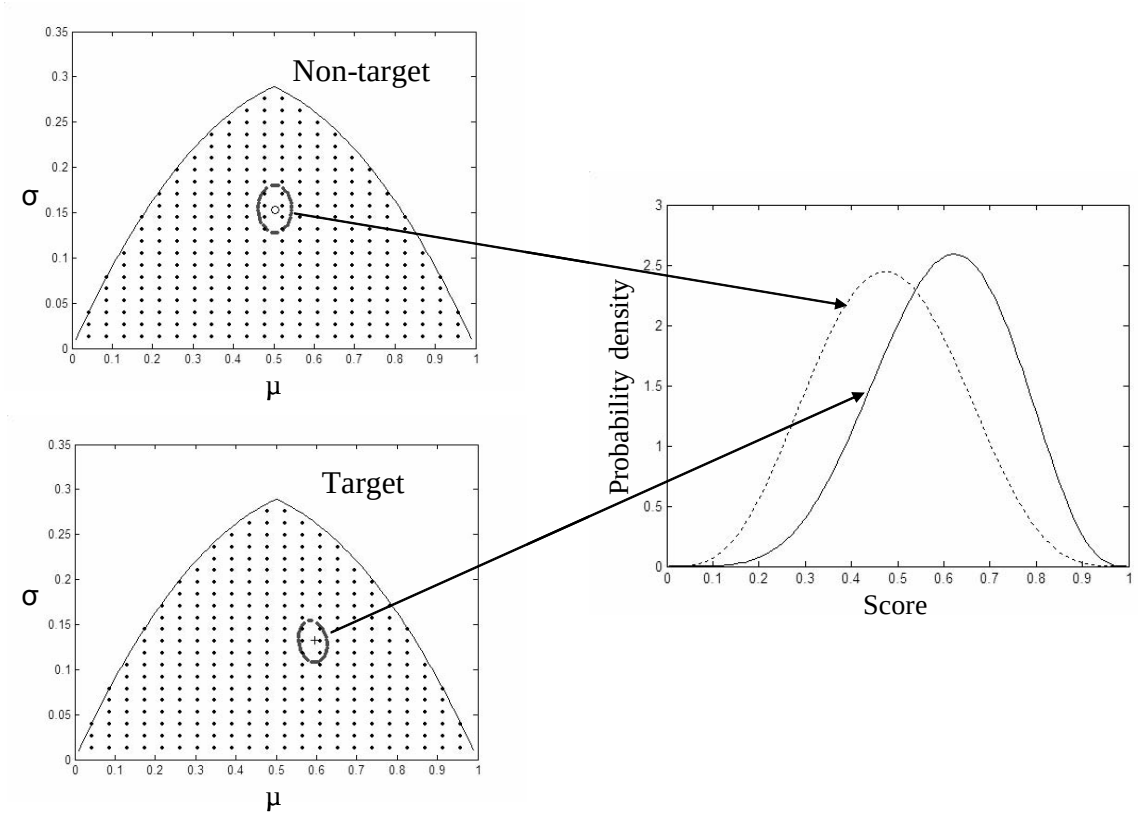


Figure 3.6 Bayesian posterior densities of parameters. The two plots at the left show the admissible domains of means and standard deviations for beta densities. Values of standard deviation outside each "rounded triangle" do not exist for these densities. The target and non-target densities shown at the right map to specific locations on the standard deviation versus mean graphs shown at the left.

parameters given a set of samples are obtained. Further, if the target and non-target samples are independent, a joint posterior weight is obtained for any combination of target and non-target densities. The application of Bayes' rule requires the specification of prior parameter densities. The typical prior density has uniform distributions of mean and standard deviation over their admissible domains.

The following discussion outlines an analytical determination of a ROC curve density. As is typical for Bayesian evaluations, the analytical results produce integrals that are not tractable to further evaluate analytically (see MacKay [MacKay, 1992a], Bishop [Bishop, 1995], Clyde [Clyde, 1999][Clyde and George, 2004], Hoeting [Hoeting *et al.*, 1999], and Jordan [Jordan *et al.*, 1999]). However, numerical evaluation is possible for the beta density model and for more complex density models (such as beta mixture models).

Throughout the analytical progression that follows, the subscripts on density (for example the subscript $u|d$ on $p_{u|d}$) are used indicate the quantities being evaluated as random variables (see discussion in Section 2.2 regarding relation of random variables and parameters).

Let $d \equiv \{s_i |\ i = 1, ..., I\}$ be a set of known independent non-target score samples, where s_i is the ith non-target score sample, and let u be the non-target density parameters. For example, for a beta density model, u may be the (μ_n, σ_n) parameters that are the allowable means (μ_n) and standard deviations (σ_n) from the admissible set. Let $p_{u|d}(u|d)$ be the conditional probability density of the non-target score parameters u given d. Then by Bayes' rule, $p_{u|d}(u|d)$ is

$$p_{u|d}(u|d) = C_o p_{d|u}(d|u) p_u(u), \tag{3.4}$$

where the constant C_o depends on d, where $p_{d|u}(d|u)$ is the conditional probability density of the samples given the parameters and $p_u(u)$ is the prior probability density of the parameters.

For a beta probability density, Equation (3.4) is

$$p_{(\mu_n,\sigma_n)|d}(\mu_n,\sigma_n|d) = C_1 p_{d|(\mu_n,\sigma_n)}(d|\mu_n,\sigma_n) p_{\mu_n,\sigma_n}(\mu_n,\sigma_n), \tag{3.5}$$

where the constant C_1 depends on d.

By sample independence, the probability density of the samples given the non-target score parameters, $p_{d|(\mu_n,\sigma_n)}(d|\mu_n,\sigma_n)$ is

$$p_{d|(\mu_n,\sigma_n)}(d|\mu_n,\sigma_n) = C_2 \prod_{i=1}^{I} \frac{s_i^{\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1]-1}(1-s_i)^{\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1][\frac{1}{\mu_n}-1]-1}}{\frac{\Gamma(\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1])\Gamma(\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1][\frac{1}{\mu_n}-1])}{\Gamma(\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1]+\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1][\frac{1}{\mu_n}-1])}}, \tag{3.6}$$

where the constant C_2 depends on d.

Thus,

$$\begin{aligned} &p_{(\mu_n,\sigma_n)|d}(\mu_n,\sigma_n|d) \\ &= C_3\{\prod_{i=1}^{I} \frac{s_i^{\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1]-1}(1-s_i)^{\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1][\frac{1}{\mu_n}-1]-1}}{\frac{\Gamma(\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1])\Gamma(\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1][\frac{1}{\mu_n}-1])}{\Gamma(\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1]+\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1][\frac{1}{\mu_n}-1])}}\} p_{\mu_n,\sigma_n}(\mu_n,\sigma_n), \end{aligned} \tag{3.7}$$

where the constant C_3 depends on d.

If the assumption is made that $p_{\mu_n,\sigma_n}(\mu_n, \sigma_n)$ is uniform over all allowable values of μ_n, σ_n, then

$$p_{(\mu_n,\sigma_n)|d}(\mu_n, \sigma_n|d) = C_4 \prod_{i=1}^{I} \frac{s_i^{\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1]-1}(1-s_i)^{\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1][\frac{1}{\mu_n}-1]-1}}{\frac{\Gamma(\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1])\Gamma(\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1][\frac{1}{\mu_n}-1])}{\Gamma(\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1]+\mu_n[\frac{\mu_n(1-\mu_n)}{\sigma_n}-1][\frac{1}{\mu_n}-1])}}, \qquad (3.8)$$

where the constant C_4 depends on d.

The points (μ_n, σ_n) chosen within the admissible set are used to estimate Bayesian posterior densities. Each Bayesian posterior density may be visualized as the three-dimensional function described by Equation (3.8) that is non-zero for any value within the admissible set. The uniformly spaced points shown in the plots on the left in Figure 3.6 select the elements of u and v that are evaluated numerically.

Let $h \equiv \{q_j |\ j = 1, ..., J\}$ be a set of known independent target score samples, where q_j is the jth target score sample, and let v be the target density parameters. For example, for a beta density model, v may be the (μ_t, σ_t) parameters that are the allowable means (μ_t) and standard deviations (μ_t) from the admissible set. Then applying the analysis above yields expressions similar to Equations (3.5) to (3.8), where the expression for $p_{\mu_t,\sigma_t|h}(\mu_t, \sigma_t|h)$ is obtained by replacing i with j, I with J, u with v, and (μ_n, σ_n) with (μ_t, σ_t) in Equation (3.8).

Theorem 3.1 Posterior density evaluation for the parameters given the non-target samples

Let $p_{s|u}(s_i|u_k)$ be the non-target score probability density evaluated at the ith non-target score sample given the kth non-target sample parameter u_k, where u_k specifies a vector. Let $p_{u|d}(u_k|d)$ be the probability density of the non-target sample parameters evaluated at u_k given the non-target samples d, where $d \equiv \{s_{_i}|\ i = 1, ..., I\}$. Let $p_u(u_k)$ be the prior probability density of the non-target sample parameter vector evaluated at u_k. Assume that the non-target samples are independent and identically distributed. Then

$$p_{u|d}(u_k|d) = C_5 \prod_{i=1}^{I} p_{s|u}(s_{_i}|u_k) p_u(u_k), \tag{3.9}$$

where the constant C_5 depends on d.

Proof

By non-target sample independence and identical distribution

$$p_{d|u}(d|u_k) = C_6 \prod_{i=1}^{I} p_{s|u}(s_{_i}|u_k), \tag{3.10}$$

where the constant C_6 depends on d.

From Bayes' rule

$$p_{u|d}(u_k|d) = C_7 p_{d|u}(d|u_k) p_u(u_k), \tag{3.11}$$

where the constant C_7 depends on d.

Therefore, combining Equations (3.10) and (3.11) yields (3.9).

As an example, for a beta density model, u_k specifies a mean and standard deviation. An expression for $p_{v|h}(v_m|h)$ is developed similarly.

In Figure 3.6, the oval regions shown in the vicinity of the target and non-target mean and standard deviation values provide a confidence contour for the posterior probability that the given set of samples is obtained from densities parameterized by the indicated regions. An example of a graph of Bayesian posterior density is shown in Figure 3.7. A plane that intersects the graph of the density such that a selected percentage (e.g., 90%) of the volume of the density is enclosed defines a confidence contour.

Definition - Confidence contour for the non-target parameter density

Let $p_{u|d}(u|d)$ be the probability density of the non-target sample parameters given the non-target samples d. Let $c.c.$ be the desired confidence coverage (e.g., if the desired coverage is 90%, then the confidence contour fraction is 0.90). Let u have elements (μ_n, σ_n) in the domain of the admissible set. For any $z \geq 0$, let N_z consist of the set of all (μ_n, σ_n) where $p_{(\mu_n,\sigma_n)|d}(\mu_n, \sigma_n|d) \geq z$.

$$\widehat{z} = \max \left\{ z \geq 0 : \iint\limits_{N_z} p_{(\mu_n,\sigma_n)|d}(\mu_n, \sigma_n|d) d\sigma_n d\mu_n \geq c.c. \right\} \tag{3.12}$$

N_z is the the set of u (within the admissible set) that provides the desired confidence coverage ($c.c$).

To evaluate numerically, let

$$\widehat{z}_{test} = \max_{\mathcal{A}} (p_{(\mu_n,\sigma_n)|d}(\mu_n, \sigma_n|d)). \tag{3.13}$$

Find $N_{\widehat{z}_{test}}$ for $\widehat{z}_{test}$. Then find $c.c.test$ for$N_{\widehat{z}_{test}}$. If $c.c.test < c.c.$, then let $\widehat{z}_{test} = \widehat{z}_{test} - \varepsilon$. The value ε is a selected step size by which the change in the value of

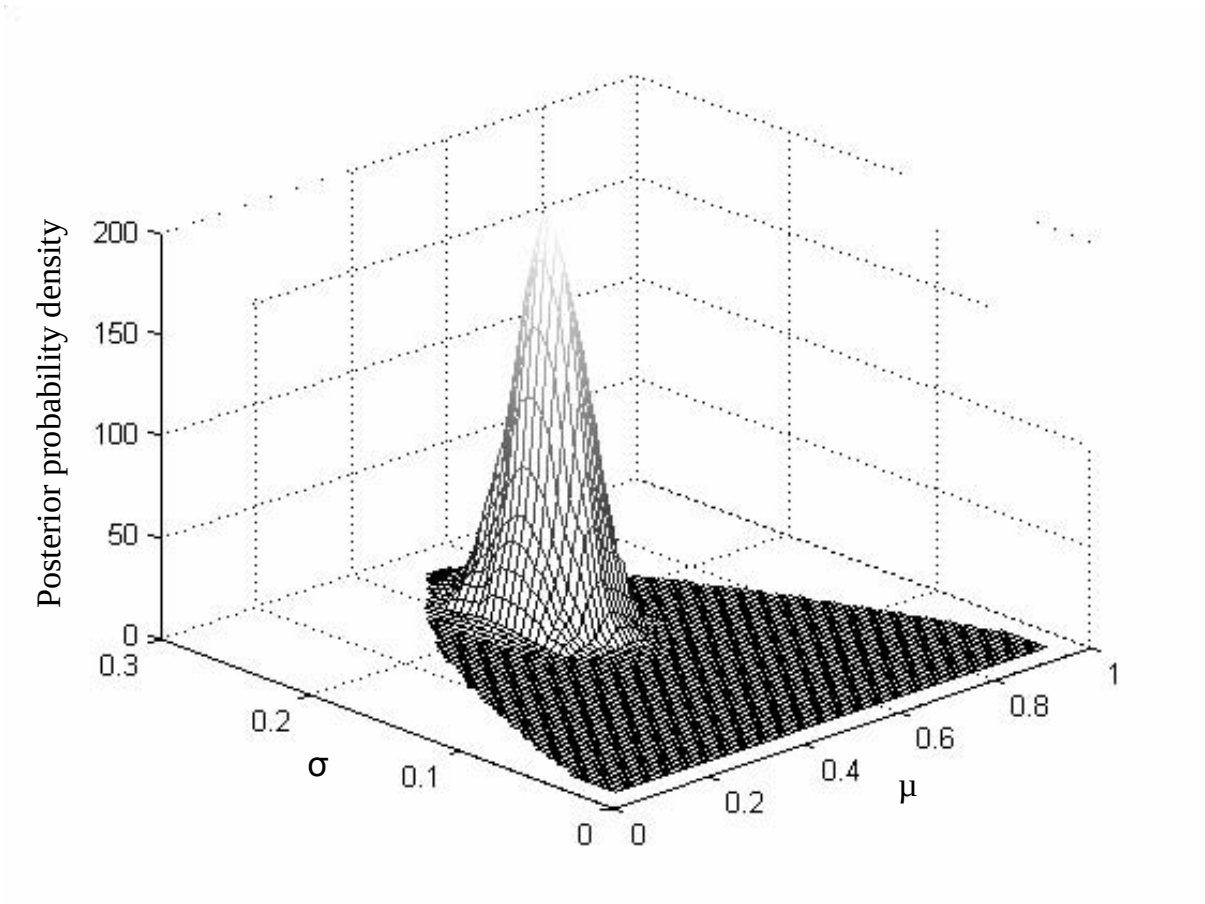


Figure 3.7 Bayesian posterior density of beta density parameters. The posterior density formed from 300 mean μ and standard deviation σ pairs with respect to a set of 30 target samples from a beta density of score is shown (a similar plot applies for 30 non-target samples). The maximum likelihood estimate for the mean and standard deviation is at the peak of the displayed density.

$\widehat{z}_{test}$ is specified. Repeat the process, continuing to reduce $\widehat{z}_{test}$ until $c.c.test = c.c.$. The confidence contour for the target parameter density is developed similarly.

By Bayes' rule and assumed sample independence, the posterior probability that a selected target mean and standard deviation are the parameters that specify the true (or underlying) beta target density given a set of samples is proportional to the product of all density values for the samples multiplied by the prior probability density of the parameters. This process of evaluating the posterior density is repeated for a set of non-target samples. Then the results are multiplied to obtain a value proportional to the probability that a pair of target parameters and a pair of non-target parameters are the parameters of the underlying target and non-target densities of scores. The posterior density in Figure 3.7 illustrates Equation (3.9). Any point within the admissible set is weighted by

$$w_k = \prod_{i=1}^{I} p_{s|u}(s_i|u_k)p_u(u_k), \tag{3.14}$$

where w_k is the weight for the non-target parameters u_k. A similar expression applies for $w_{m,}$ where w_m is the weight for point v_m and the replacement of k by m indicates target point m.

Let the product $w_k w_m$ be the combined posterior weighting of a target and non-target density pair (evaluated at $u_{k,}\, v_m$). From Equation (3.14) for w_k and the similar expression for w_m,

$$w_k w_m = \prod_{k=1}^{K} p_{d|u}(d|u_k)p_u(u_k) \prod_{m=1}^{M} p_{h|v}(h|v_m)p_v(v_m). \tag{3.15}$$

From Equation (2.5), $\widehat{F_k}(t;u_k) = \int_t^\infty f(s;u_k)ds,$ and from Equation (2.6), v_m is $\widehat{G}(t;v_m) = \int_t^\infty g(s;v_m)ds$. Thus, from Equation (2.10) the ROC curve is

$$r_{k,m}(x;u_{k,}v_m) = \widehat{G}(\widehat{F}^{-1}(x;u_k), v_m). \tag{3.16}$$

Theorem 3.2 ROC curve density

Let d =$\{s_1 ... s_I\}$ be a set of independent and identically distributed samples s_i from distribution f and let h =$\{q_1 ... q_J\}$be a set of independent and identically distributed samples q_j from distribution g, where $i = 1, 2, ..., I$ and $j = 1, 2, ..., J$. Let $p_u(u)$ and $p_v(v)$ be prior densities of the random parameter vectors u and v. Let $p_{y|x}(y|x, d, h)$ be the probability density of correct detection probability y given false alarm probability x and d and h. Then

$$p_{y|x}(y|x,d,h) = C_8 \iint_{\mathcal{A}} p_{y|x}(y|x,u,v) \prod_{i=1}^{I} f(s_i|u) p_u(u) \prod_{j=1}^{J} g(q_j|v) p_v(v) du dv, \quad (3.17)$$

where the contant C_8 depends on d and h and the limits of integration are over the admissible set $\mathcal{A}$.

Proof. See Appendix A-2.

Substituting the beta density parameters and admissible set into Equation (3.17):

$$p_{y|x}(y|x,d,h) = C_9 \int_0^{.5} \int_0^{\frac{1-\mu_t}{\mu_t(\mu_t+2)(\mu_t+1)^2}} \int_0^{.5} \int_0^{\frac{1-\mu_n}{\mu_n(\mu_n+2)(\mu_n+1)^2}} p_{y|x}(y|x,u,v)$$

$$\cdot \prod_{i=1}^{I} f(s_i;u) \prod_{j=1}^{J} g(q_j;v) p_u(u) p_v(v) d\sigma_n d\mu_n d\sigma_t d\mu_t$$

$$+ \int_{.5}^{1} \int_0^{\frac{\mu_t(1-\mu_t)^2}{2-\mu_t}} \int_{.5}^{1} \int_0^{\frac{\mu_n(1-\mu_n)^2}{2-\mu_n}} p_{y|x}(y|x,u,v)$$

$$\cdot \prod_{i=1}^{I} f(s_i;u) \prod_{j=1}^{J} g(q_j;v) p_u(u) p_v(v) d\sigma_n d\mu_n d\sigma_t d\mu_t$$

$$+ \int_0^{.5} \int_0^{\frac{1-\mu_n}{\mu_n(\mu_n+2)(\mu_n+1)^2}} \int_{.5}^{1} \int_0^{\frac{\mu_n(1-\mu_n)^2}{2-\mu_n}} p_{y|x}(y|x,u,v)$$

$$\cdot \prod_{i=1}^{I} f(s_i|u) \prod_{j=1}^{J} g(q_j|v) p_u(u) p_v(v) d\sigma_n d\mu_n d\sigma_t d\mu_t$$

$$+ \int_{.5}^{1} \int_0^{\frac{\mu_t(1-\mu_t)^2}{2-\mu_t}} \int_0^{.5} \int_0^{\frac{1-\mu_n}{\mu_n(\mu_n+2)(\mu_n+1)^2}} p_{y|x}(y|x,u,v)$$

$$\cdot \prod_{i=1}^{I} f(s_i|u) \prod_{j=1}^{J} g(q_j|v) p_u(u) p_v(v) d\sigma_n d\mu_n d\sigma_t d\mu_t, \tag{3.18}$$

where

$$p_{y|x}(y|x,u,v) = p_{y|x}(y|x,\mu_n,\sigma_n,\mu_{t,}\sigma_t) \tag{3.19}$$

$$\prod_{i=1}^{I} f(s_i|u) = \prod_{i=1}^{I} f(s_i|\mu_n,\sigma_n) \tag{3.20}$$

$$\prod_{j=1}^{J} g(q_j|v) = \prod_{j=1}^{J} g(q_j|\mu_t,\sigma_t) \tag{3.21}$$

$$p_u(u) = p_{\mu_{n,}\sigma_n}(\mu_n,\sigma_n) \tag{3.22}$$

$$p_v(v) = p_{\mu_t,\sigma_t}(\mu_t, \sigma_t), \tag{3.23}$$

where the contant C_9 depends on d and h.

Lemma 3.1 Discretization of posterior densities

Let d be a set of independent and identically distributed samples s_i of f and let h be a set of independent and identically distributed samples q_j of g, where $i = 1, 2, ..., I$ and $j = 1, 2, ..., J$. Let $p_u(u)$ and $p_v(v)$ be prior densities of the parameter vectors u and v with elements (μ_n, σ_n) and (μ_t, σ_t), respectively. Let u_k and v_m be u and v selected uniformly over the parameter domains within the admissible set. Finally, let $A_k = (\mu_{n,(k+1)} - \mu_{n,k})(\sigma_{n,(k+1)} - \sigma_{n,k})$ and $\Delta_n = (\mu_{n,(k+1)} - \mu_{n,k})(\sigma_{n,(k+1)} - \sigma_{n,k})$ and let $A_m = (\mu_{t,(m+1)} - \mu_{t,m})(\sigma_{t,(m+1)} - \sigma_{t,m})$ and $\Delta_t = (\mu_{t,(m+1)} - \mu_{t,m})(\sigma_{t,(m+1)} - \sigma_{t,m})$, where the second subscript designates position in the admissible set domain.

Then

$$C_{10} \int_{\mathcal{A}} \int_{\mathcal{A}} \prod_{i=1}^{I} [f(s_i|\mu_n, \sigma_n) p_{\mu_n,\sigma_n}(\mu_n, \sigma_n)] d\sigma_n d\mu_n$$

$$= C_{11} \lim_{K\to\infty} \sum_{k=1}^{K} \prod_{i=1}^{I} p_{s|\mu_n,\sigma_n}(s_i|\mu_{n,k}, \sigma_{n,k}) p_{\mu_n,\sigma_n}(\mu_{n,k}, \sigma_{n,k}) \tag{3.24}$$

and

$$C_{12} \int_{\mathcal{A}} \int_{\mathcal{A}} \prod_{j=1}^{J} [g(q_j|\mu_t, \sigma_t) p_{\mu_t,\sigma_t}(\mu_t, \sigma_t)] d\sigma_t d\mu_t$$

$$= C_{13} \lim_{M \to \infty} \sum_{m=1}^{M} \prod_{j=1}^{J} p_{s|\mu_t,\sigma_t}(q_j|\mu_{t,m}, \sigma_{t,m}) p_{\mu_{t,}\sigma_t}(\mu_{t,m}, \sigma_{t,m}), \tag{3.25}$$

where the constant C_{10} depends on d, the constant C_{11} depends on C_{10} and A_k, the constant C_{12} depends on h, the constant C_{13} depends on C_{12} and A_m, and the the limits of integration are over the admissible set $\mathcal{A}$.

Proof

Since each evaluated ($\mu_{n,k}$, $\sigma_{n,k}$) is uniformly spaced on the admissible set, $K \propto 1/\Delta_n$ and $M \propto 1/\Delta_t$, then the lemma follows by definition of a double integral and by limit of a Riemann sum (see [Larson *et al.*, 2002]).

Theorem 3.3 Numerical approximation of ROC curve density

Let $p_{y|x}(y|x,d,h)$ be the density of correct detection probability y given false alarm probability x and d and h. Let $p_{y|x}(y|x) = \delta(y - r(x;w))$, where δ is the dirac density or distribution function, let $p_u(u)$ be the prior density of the non-target parameters,and $p_v(v)$ be the prior density of the target parameters. Let $\{(u_k, v_m) : k = 1, ..., K, m = 1, ..., M\}$ be uniformly selected over the admissible set of u and v for the target and non-target parameter densities. Let $p_{s|u}(s_i|u_k)$ be the density of the independent and identically distributed non-target samples evaluated at the ith non-target sample s_i given the kth non-target sample parameters u_k, where u_k has elements (μ_k, σ_k) over the admissible set. Let $p_{u|d}(u|d)$ be the density of the non-target sample parameters given the non-target samples d. Let $p_u(u_k)$ be the prior density of the non-target sample parameter vector evaluated at u_k, and let $f(s_i|\mu_{n,k}, \sigma_{n,k}) = p_{s|(\mu_n,\sigma_n)}(s_i|\mu_{n,k}, \sigma_{n,k})$. Let $p_{s|v}(q_j|v_m)$ be the density of the independent and identically distributed target samples evaluated at the jth target sample q_j given the mth target sample parameters v_m, where v_m has elements $(\mu_{t,k}, \sigma_{t,k})$ over the admissible set. Let $p_{v|h}(v|h)$ be the density of the target sample parameters given the target samples h. Let $p_v(v_m)$ be the prior density of the target sample parameter vector evaluated at v_m, and let
$g(q_j|\mu_{t,m}, \sigma_{t,m}) = p_{s|\mu_t,\sigma_t}(q_j|\mu_{t,m}, \sigma_{t,m})$. Finally, let

$$\gamma(d) = \lim_{K\to\infty} \sum_{k=1}^{K} \prod_{i=1}^{I} [p_{s|(\mu_n,\sigma_n)}(s_i|\mu_{n,k}, \sigma_{n,k}) p_{\mu_n,\sigma_n}(\mu_{n,k}, \sigma_{n,k})] \tag{3.26}$$

and

$$\gamma'(d) = \iint_{\mathcal{A}} \prod_{i=1}^{I} [f(s_i|\mu_n, \sigma_n) p_{\mu_n,\sigma_n}(\mu_n, \sigma_n)] du, \tag{3.27}$$

where the limits of integration are over the admissible set $\mathcal{A}$.

Then

$$p_{y|x}(y|x,d,h)$$

$$= C_{14} \lim_{\substack{K\to\infty \\ M\to\infty}} \sum_{k=1}^{K}\sum_{m=1}^{M} \delta(y - r(x;u_k,v_m)) \prod_{i=1}^{I}[f(s_i|u_k)p_u(u_k)] \prod_{j=1}^{J}[g(q_j|v_m)p_v(v_k)], \quad (3.28)$$

where the constant C_{14} depends on K, M, d and h.

Proof

Since from Lemma 3.1, $\gamma'(d) \propto \gamma(d)$,

$$\gamma'(d) \iint_{\mathcal{A}} [\prod_{j=1}^{J} g(q_j|\mu_t,\sigma_t) p_{\mu_t,\sigma_t}(\mu_t,\sigma_t)] d\sigma_t d\mu_t$$

$$= C_{15}\gamma(d) \sum_{m=1}^{M} \prod_{j=1}^{J} [p_{s|(\mu_t,\sigma_t)}(q_i|\mu_{t,m},\sigma_{t,m}) p_{\mu_t,\sigma_t}(\mu_{t,m},\sigma_{t,m})], \quad (3.29)$$

where the constant C_{15} depends on K, M, d and h.

$$\iint_{\mathcal{A}} \iint_{\mathcal{A}} \prod_{i=1}^{I} [f(s_i|\mu_n,\sigma_n) p_{\mu_n,\sigma_n}(\mu_n,\sigma_n)] \prod_{j=1}^{J} [g(q_j|\mu_t,\sigma_t) p_{\mu_t,\sigma_t}(\mu_t,\sigma_t)] d\sigma_n d\mu_n d\sigma_t d\mu_t$$

$$= C_{16} \sum_{k=1}^{K}\sum_{m=1}^{M} \prod_{i=1}^{I} [p_{s|(\mu_n,\sigma_n)}(s_i|\mu_{n,k},\sigma_{n,k}) p_{\mu_n,\sigma_n}(\mu_{n,k},\sigma_{n,k})]$$

$$\cdot \prod_{j=1}^{J} [p_{s|(\mu_t,\sigma_t)}(q_i|\mu_{t,m},\sigma_{t,m}) p_{\mu_t,\sigma_t}(\mu_{t,m},\sigma_{t,m})], \quad (3.30)$$

where the constant C_{16} depends on on K, M, d and h.

Thus,

$$\iint_{\mathcal{A}} \iint_{\mathcal{A}} p_{y|x}(y|x,u,v) \prod_{i=1}^{I} [f(s_i|\mu_n,\sigma_n) p_{\mu_n,\sigma_n}(\mu_n,\sigma_n)]$$

$$\cdot \prod_{j=1}^{J} [g(q_j|\mu_t,\sigma_t) p_{\mu_t,\sigma_t}(\mu_t,\sigma_t)] d\sigma_n d\mu_n d\sigma_t d\mu_t \tag{3.31}$$

$$= C_{17} \sum_{k=1}^{K} \sum_{m=1}^{M} p_{y|x}(y|x,u,v) \prod_{i=1}^{I} [p_{s|(\mu_n,\sigma_n)}(s_i|\mu_{n,k},\sigma_{n,k}) p_{\mu_n,\sigma_n}(\mu_{n,k},\sigma_{n,k})]$$

$$\cdot \prod_{j=1}^{J} [p_{s|\mu_t,\sigma_t}(q_i|\mu_{t,m},\sigma_{t,m}) p_{\mu_t,\sigma_t}(\mu_{t,m},\sigma_{t,m})], \tag{3.32}$$

where the constant C_{17} depends on K, M, d and h.

The theorem follows upon substituting Equation (3.32) into (3.31) and using Equation (3.17).

To extend the above theorem to the CEG curve, let the CEG curve be defined as (see Section 2.3)

$$P(T|s,u_k,v_m) = \frac{g(s|T,v_m)P(T)}{g(s|T,v_m)P(T) + f(s|N,u_k)P(N)}, \tag{3.33}$$

where $s \in [0,1]$. Let $\{(u_k,v_m) : k = 1,...,K, m = 1,...,M\}$ be uniformly selected over the admissible set of u and v for the target and non-target parameter densities. Let $\widetilde{y}$ denote a selected location on the vertical axis of the CEG curve (see Figure 2.2 for a CEG curve plot). Let $P(T|s,u_k,v_m)$ be the probability of target event given score, u_k and v_m, let $g(s|T,v_m)$ be the density of score given target event and v_m, let $f(s|N,u_k)$ be the probability density of score given non-target event and u_k, let $P(T)$ be the prior probability of target event, and let $P(N)$ be the prior probability of non-target event. Replace $r(x;u_k,v_m)$ by $P(T|s,u_k,v_m)$. Then the probability density of the probability of

target given score for any evaluated score value is

$$p_{P(T|s)}(P(T|s), d, h)$$

$$= C_{15} \lim_{\substack{K\to\infty \\ M\to\infty}} \sum_{k=1}^{K} \sum_{m=1}^{M} \delta(\widetilde{y} - P(T|s; u_k, v_m)) \prod_{i=1}^{I} [f(s_i|u_k) p_u(u_k)] \prod_{j=1}^{J} [g(q_j|v_m) p_v(v_k)], \tag{3.34}$$

where the constant C_{15} depends on K, M, d and h.

Note that covering the entire admissible parameter space volume with a practical number of grid points becomes computationally more difficult as the number of dimensions increases (see [Gelman *et al.*, 2004]). For higher dimensions, Monte Carlo methods (see [Hammersley and Handscomb, 1964], [Kass and Raftery, 1995]) or related approximation methods may be used (such as Gibbs sampling or the Metropolis Algorithm; see [Casella and Berger, 2002], [MacKay, 2003]); where i.i.d. sampling assumptions are necessary.

Note that a fundamental assumption for a simple Monte Carlo approach (see [Hammersley and Handscomb, 1964] and [Kass and Raftery, 1995]) is

$$\int p_s(s|u) p_u(u) du = C_{16} \lim_{k\to\infty} \sum_{k=1}^{K} p_s(s|u_k) p_u(u_k), \tag{3.35}$$

where the constant C_{16} depends on K,and the K grid points are independently and identically selected from the admissible set. Equation (3.35) may replace Equation (3.24) for i.i.d. sampling rather than uniform grid selection; thus the framework described here is appropriate for Monte Carlo methods.

Calculating $w_{k,m}$ values and the $r_{k,m}(x; u_{k,}v_m)$ function is straightforward and numerically tractable. However, it is desirable to limit the size of K and M by removing the regions where w_{km} approach zero (i.e., select only u_k and v_m values such that w_{km} is

greater than a given small value). For computational efficiency, an iterative process is used. The iterative process is described below; Section 5.5 gives a full description of the numerical evaluation process used for the results shown in Chapters 4 and 5.

Procedure 3.1 Iterative Process for calculating weight values

1. Select K evaluation points $k = 1, 2, ..., K$ (e.g., $K = 300$) that are uniform over the admissible set of non-target score parameters $u_{k,}$ where each u_k consists of mean $\mu_{n,k}$ and standard deviation $\sigma_{n,k}$.

2. Select M evaluation points $m = 1, 2, ..., M$ (e.g., $M = 300$) that are uniform over the admissible set of target score parameters $v_{m,}$ where each v_m consists of mean $\mu_{t,m}$ and standard deviation $\sigma_{t,m}$.

3. Find $w_k = \prod_{i=1}^{I}[p_{s|u}(s_i|u_k)p_u(u_k)]$ for each evaluation point selected in step 1 and for a given set of I target samples $s_i, i = 1, 2, ..., I$.

4. Find $w_m = \prod_{j=1}^{J}[p_{s|v}(s_j|v_m)p_v(v_m)]$ for each evaluation point selected in step 2 and for a given set of J target samples $s_j, j = 1, 2, ..., J$.

5. Combine all w_k and w_m pairs from steps 3 and 4 to find the initial values (e.g., 90,000) of $w_k w_m$.

6. Find the root mean squared distance to the mean of the parameter values for each $(\mu_{n,k,}\sigma_{n,k})$ pair, i.e., $[(\mu_{n,k} - \frac{1}{K}\sum_{k=1}^{K}\mu_{n,k})^2 + (\sigma_{n,k} - \frac{1}{K}\sum_{k=1}^{K}\sigma_{n,k})^2]^{1/2}$.

7. Repeat step 6 for each $(\mu_{t,m,}\sigma_{t,m})$ pair.

8. Retain a subset of the combinations of the $w_k w_m$ pairs that are closest in distance as defined by steps 6 and 7 to the mean of non-target and target parameter values, respectively. Also, retain any additional w_k and w_m pairs without regard to distance

whose $w_k w_m$ value is greater than the lowest $w_k w_m$ value of the subset of pairs that are closest in distance.

9. Create a new uniform grid of target mean and standard deviation values (for example, 10 x 10) that bound the region formed by the pairs retained in steps 6 and 8; the new grid forms new $(\mu_{n,k,}\sigma_{n,k})$ pairs (for example 100).

10. Create as in step 9 a new uniform grid of non-target means and standard deviation values (for example, 10 x 10) that bound the region formed by the pairs identified in steps 7 and 8; the new grid forms new $(\mu_{t,m,}\sigma_{t,m})$ pairs (for example, 100).

11. Find the posterior weightings $w_k w_m$ of the new pairs (e.g., 10,000 posterior weightings).

12. Retain all (w_k, w_m) pairs such that 99.9% of the total posterior parameter weightings are maintained.

13. Repeat steps 9 through 12, except use the region formed by the pairs identified in step 12 rather than step 9.

As the number of non-target samples and target samples increases, the probability density shown in Figure 3.7 is more highly peaked, and the region where the weights w_k and w_m have significant magnitudes is smaller.

Theorem 3.4 True versus possible parameter sets

Let $d = \{s_i : i = 1, ..., I\} \subset [0, 1]$ be a set of independent and identically distributed samples s_i of the density of non-target samples $f(s; u)$. Let $F_I(s)$ be the distribution of these samples. Let $u_{\widetilde{b}}$ be the true (underlying) parameter values of the non-target density f. Let u_z be a possible parameter from the admissible set $\mathcal{A}$, and let

$$c_z = \prod_{i=1}^{I} f(s_i; u_{\widetilde{b}}) - \prod_{i=1}^{I} f(s_i; u_z). \tag{3.36}$$

Then, as $I \to \infty$, c_z increases for all $z \neq \widetilde{b}$.

Proof

By definition of independent and identically distributed samples, the distribution of the samples $F_I(s)$ equals the distribution of the random variable S (see [Papoulis, 1991, pp. 185]) as $I \to \infty$ (see [Stark and Woods, 1986, pp. 252]). Thus, as $I \to \infty$, c_z increases for all $z \neq \widetilde{b}$.

A similar result holds true for the target samples. Further, since the ROC curve density combines the target and non-target posterior densities (see Equation (3.17)), the ROC curve density also narrows (the ROC curve density evaluated at a given false alarm probability approaches a dirac distribution) as sample size increases.

Figure 3.8 shows the final step in the generation of the ROC curve density. Based on the posterior density calculations for the target and non-target parameters (i.e., the mean and standard deviation for a beta density), an approximation of the ROC curve density is developed. A selected target density of score, a selected non-target density of score, and a varying threshold forms a ROC curve and has a weight. Many sets of selections result in many ROC curves, each with a weight $w_k w_m$. The figure shows curves that represent $\delta[y - r_{k,m}(x; u_{k,}v_m)]$ for five selected k and m pairs. The weighted summation of

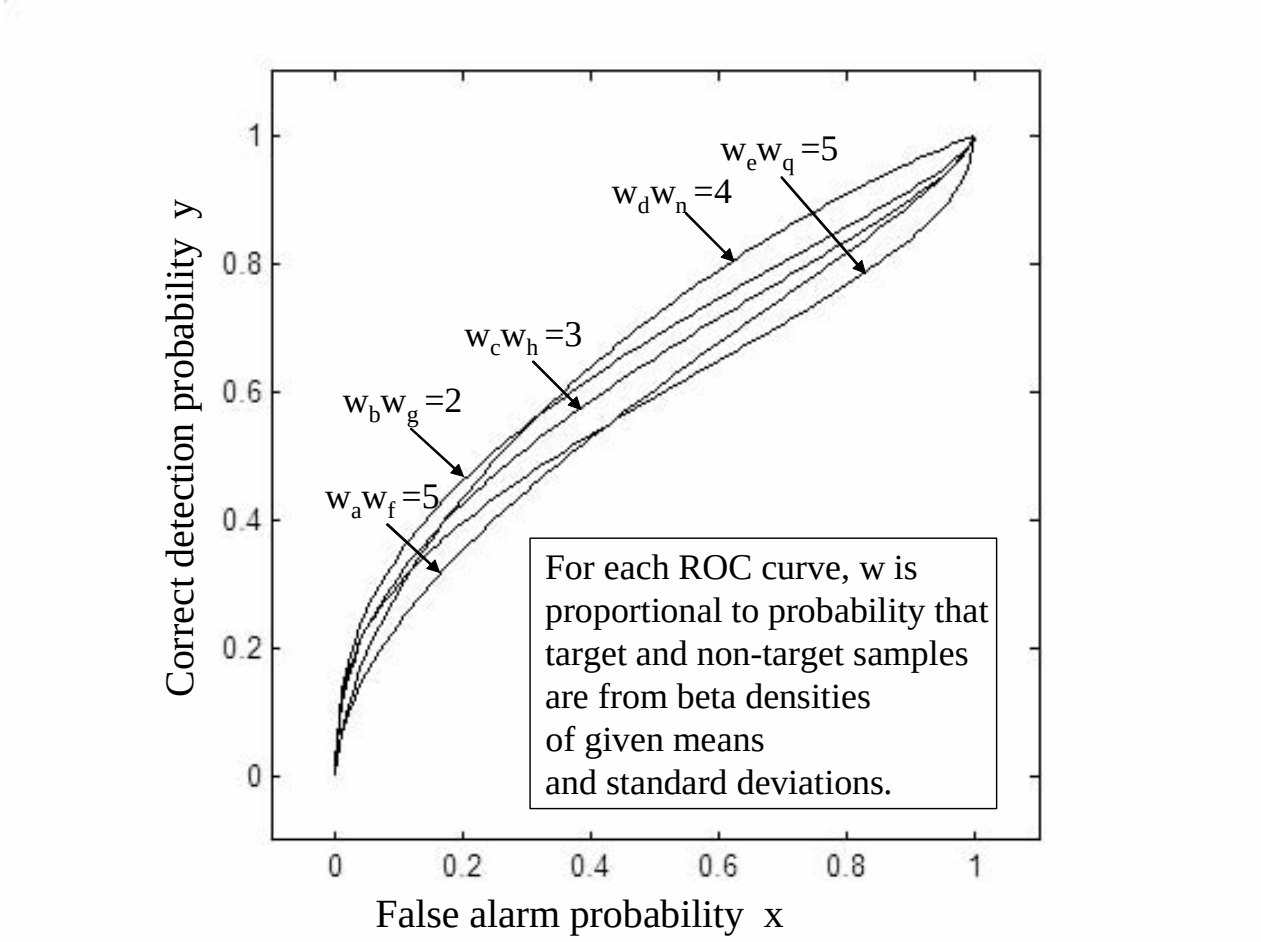


Figure 3.8 Weighted ROC curves. Based on the posterior density approximations for the target and non-target parameters values (i.e., the mean and standard deviation for a beta density), an approximation of the ROC curve density is developed. The combination of a selected target density of score and a selected non-target density of score forms a ROC curve and has a weight. Many sets of selections results in many ROC curves, each with a weight $w_k w_m$. Here only five weighted ROC curves are shown; for a large number of weighted ROC curves many descriptive statistics may be computed, such as median estimates for the ROC curve, confidence intervals for the ROC curve, median estimates for the AUC value, and confidence intervals for the AUC value.

$\delta[y - r_{k,m}(x; u_{k,}v_m)]$ for k and m selected from the admissible set is described by Equation (3.28). Five ROC curves are shown; a much larger number of weighted ROC curves are needed to represent a ROC probability density model (approximately 10,000 ROC curves are typically employed). As K and M become large, these weighted curves approximate the analytical ROC surface density. In particular, if a large number of $\delta[y - r_{k,m}(x; u_{k,}v_m)]$ functions for selected k and m pairs are each replicated a number of times proportional to $w_k w_{m,}$ then the set of replicated functions represents the density of ROC curves (as the preceding theorem indicates). For a large number of weighted ROC curves, many descriptive statistics may be computed, such as median estimates for the ROC curve, confidence intervals for the ROC curve, median estimates for the AUC value, and confidence intervals for the AUC value as detailed in Chapter 3. This outcome extends in a straightforward manner to the CEG curve, and Section 4.2.5 applies the method described here to CEG curves.

The above discussion is self-contained in that an analytical ROC curve density process is developed. Necessary inputs include non-target and target samples, specified density models for the target and non-target samples, and prior densities for the parameters of the models. The selection of evaluation points for the prior densities enables a numerical estimate of the ROC curve density.

The upper left plot of Figure 3.9 shows selected target parameter points (circles) and the upper right plot shows example non-target parameter points (circles). The lower left plot shows target densities (solid curves) and non-target densities (dashed curves) for these points, and the lower right plot shows the ROC curves formed by combinations of these curves: out of the 64 possible pairs, the 44 are chosen that have the highest posterior parameter density. The plots demonstrate that a slight shift in parameter value impacts density shape and the corresponding ROC curve. As increasing numbers of target and non-target samples are drawn, the densities that fit the samples well using Bayesian posterior density evaluation converge. Since a sequence of random variables converges

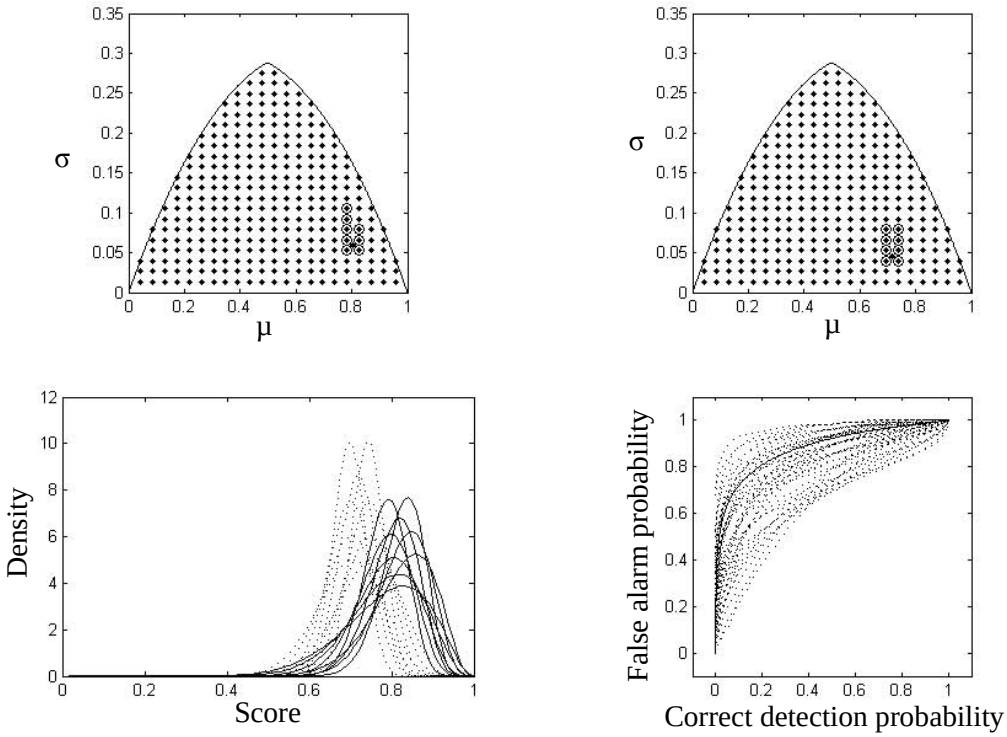


Figure 3.9 Parameter variation with corresponding densities and ROC curves. The upper left plot shows parameter points that select target densities, and the upper right plot shows parameter points that select non-target densities. The lower left plot shows target (solid curves) and non-target (dashed curves) densities for these points, and the lower right plot shows the corresponding ROC curves.

in distribution as the number of samples becomes large (see Definition 5.5.10 of [Casella and Berger, 2002, pp. 235]) assuming that the samples are i.i.d, the range of densities that have high likelihoods (i.e., that fit the samples well) narrows as sample size increases. An increase in sample size is observed experimentally to enable large regions of standard deviations and means to be disregarded, because the corresponding posterior density regions have low magnitude (see the above theorem).

Note that parameter evaluation points uniformly spaced for one parameter choice may not be uniformly spaced for other parameter choices. Figure 3.10 plots points uniformly spaced over variance and mean rather than standard deviation and mean, and then converts these points to standard deviation and mean. Comparison with Figure 3.6 shows that these points are now more concentrated at larger standard deviations. Figure 3.11 examines posterior probability density over the beta density parameters a and b rather than mean and standard deviation. As a and b increase, density width generally decreases, which initially provides better fit to samples for selected means, until a maximum posterior parameter weight is reached, beyond which the target and non-target densities have variance too small to adequately fit the samples. Thus, selecting points uniformly over a and b requires different prior assumptions than selecting points uniformly over mean and standard deviation.

In this chapter, performance metric probability densities have been developed; Chapter 4 leverages these densities to obtain and verify confidence intervals and other descriptive statistics.

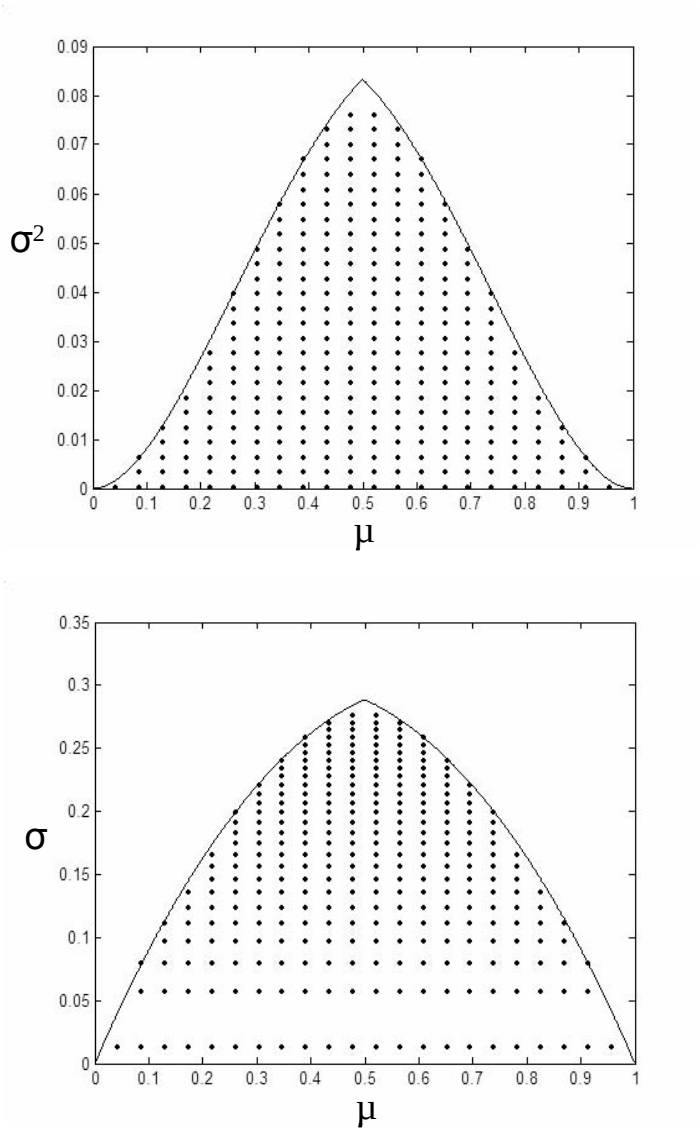


Figure 3.10 Uniformly spaced parameter selection over variance and mean compared with selection over standard deviation and mean. The curves in both plots enclose allowed beta density parameters. The points that are uniformly spaced in variance and mean are transferred to standard deviation versus mean in the lower plot. Note that while the curves are of different shape, the limits of σ and σ^2 are both defined by the admissable set of Equation (3.3) (the difference in shape is simply a result of using a vertical axis of σ rather than σ^2).

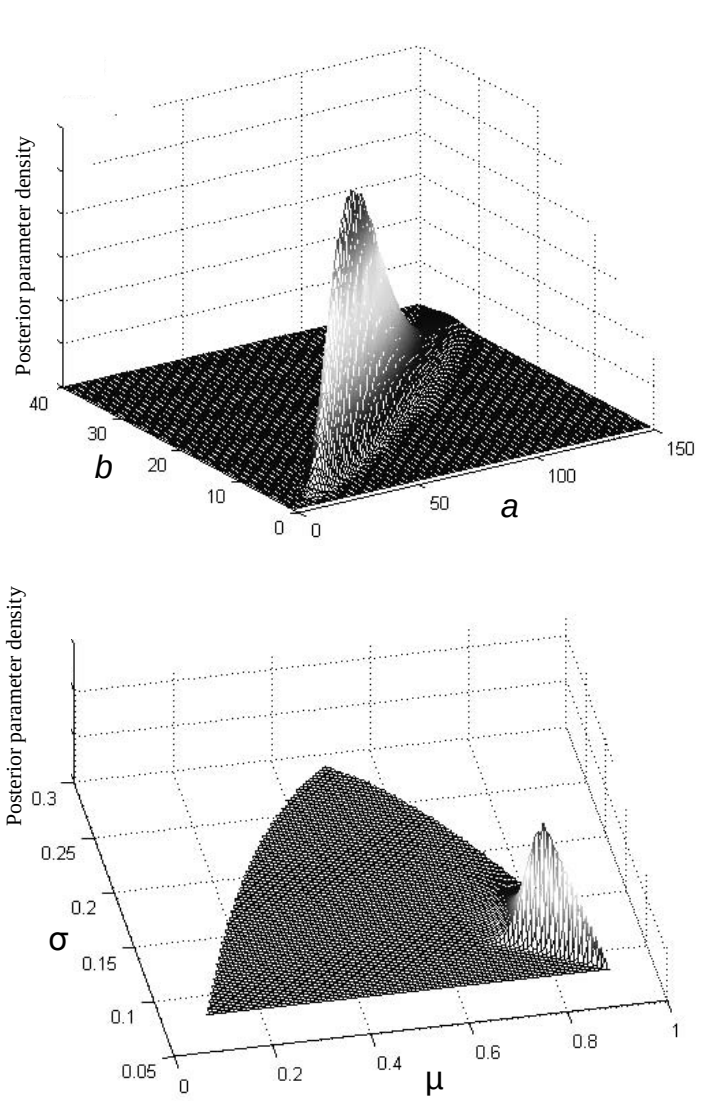


Figure 3.11 Beta posterior parameter densities that compare a and b versus σ and μ parameters. The bottom plot is as in Figure 3.7 but for a different set of target and non-target samples. The top plot shows that as a and b increase, the density width generally decreases, which initially provides better fit to samples for selected means, until a maximum posterior parameter weight is reached (here at $a = 55, b = 15$), beyond which the target and non-target densities have variance too small to adequately fit the samples.

4. Probability Density Characterization and Verification

The method of Chapter 2 generates densities for detection system performance metric curves, such as the ROC curve. Various descriptive statistics then characterize these densities; examples of such statistics are confidence contours for the ROC curve and confidence interval limits for the AUC value. Following the development of such characterization methods, a Monte Carlo approach estimates their accuracy using various examples. Coverage accuracy and alpha are used to test whether or not the defined confidence interval limits are accurate over a large number of trials. For example, suppose that 30 target samples and 30 non-target samples generate a ROC curve. Then, based on only these 60 samples, a ROC curve probability density and 90% confidence intervals can be developed. The 90% confidence intervals are intended to enclose the true ROC curve 90% of the time. This outcome can be tested by generating 30 target samples and 30 non-target samples many times, producing confidence intervals for each run, and calculating the percentage of runs in which the confidence intervals enclose truth. The coverage accuracy and alpha metrics are of particular interest because they provide quantitative means to compare the method developed here with methods in the literature.

4.1 Development of descriptive statistics

4.1.1 The AUC value densities and confidence intervals. The following process maps the weighted ROC curves shown in Figure 3.8 to AUC value uncertainty. Recall that if the target and non-target density parameters u_k and v_m are specified as described in Equation (3.16), then a deterministic ROC curve results. Further, a representative set of (k, m) pairs results in a representative set of ROC curves. Chapter 2 describes a process for generating such ROC curves (see Figure 3.8). First, find the ROC curve $r(x; u_{k,}v_m)$ for each selected (k, m) pair, where k and m identify one of the K parameters u_k and one of the M parameters v_m. Second, replicate each curve a number of times proportional to

its posterior parameter weighting $w_k w_m$, which is defined in Equation (3.15). Finally, calculate for each ROC curve a corresponding AUC value as

$$AUC(u_{k,}v_m) = \int_0^1 r(x; u_{k,}v_m)dx, \tag{4.1}$$

where y and x are correct detection probability and false alarm probability, respectively, of the ROC curve; that is, $y = r(x; u_{k,}v_m)$ is the ROC function.

Confidence intervals for the AUC values are developed as follows. Center an impulse probability density function at each observed AUC value. Add and normalize all impulse functions such that the result is a probability density. Denote this density $p_z(z)$, where z is the domain of possible AUC values. Begin at an AUC test value of 0 and increase until the AUC test value is found such that the integral of $p_z(z)$ from 0 to the AUC test value is 0.05. This test value is a lower 90% AUC confidence interval. Similarly, begin at an AUC test value of 1 and decrease until the AUC test value is 0.05. This test value is an upper 90% AUC confidence interval. These following equations describe the process:

$$\int_0^{test\ value_{lower}} p_z(z)dz = 0.05, \qquad \int_{test\ value_{upper}}^1 p_z(z)dz = 0.05 \tag{4.2}$$

In practice, the impulse function is obviously not practical to evaluate numerically. Instead, compute the lower AUC confidence interval by starting at an AUC test value of zero and stopping when 5% of the observed values are obtained, thereby approximating the inclusion of 5% of the total impulse functions that are used to form $p_z(z)$. Proceed similarly for the upper AUC confidence interval.

Note that a two-tail equal area approach is described here. Other approaches considered by Ross [Ross, 2003] describe alternative confidence interval definitions. Note also that a median ROC curve is generated by beginning at an AUC test value of 0, increasing the test value until the integral over the AUC value density from 0 to the test value is 0.5, and

specifying the ROC curve that corresponds to the test value as the median ROC curve ranked by AUC value. Note finally that the AUC value density is not typically symmetric, making a normal approximation approach in lieu of the above computation undesirable.

Figure 4.1 shows a histogram of AUC values, where each AUC value is weighted by its ROC curve weight as indicated in Figure 3.8. This histogram estimates the AUC value density given a set of target samples and non-target samples, assumed forms for the densities of score, assumed prior parameter densities, and specified sampling protocols. A method that generates a ROC curve 90% confidence band from AUC value densities is described in Section 4.1.2. Another method that generates a ROC curve 90% confidence band from the weighted ROC curve density without use of AUC values is described in Section 4.1.3.

4.1.2 Rank characterization of ROC curves by AUC values. The ROC curve confidence contours shown in Figure 4.2 are obtained as follows. First, find the lower and upper 90% confidence intervals for AUC value (as explained in the previous section). Next find the ROC curve closest to the lower 90% AUC confidence interval test value (see Equation (4.2)), and the ROC curve closest to the upper 90% AUC confidence interval test value. These two ROC curves form the lower and upper limits of a 90% confidence band. For the median or 50% ROC curve, find the median AUC value, and then find the ROC curve that has an AUC value closest to this median value.

Figure 4.2 provides no new information beyond that given by the ROC curve density of Figure 3.8, and, in fact, Figure 4.2, unlike Figure 3.8, does not indicate the shape of the ROC curve density (Figure 3.8 provides the entire ROC curve density, in contrast Figure 3.8 only provides confidence intervals that constitute a summary or partial description of this full ROC curve density). However, the ROC curve confidence intervals and the median ROC curve shown in Figure 4.2 are useful. For example, for a selected false

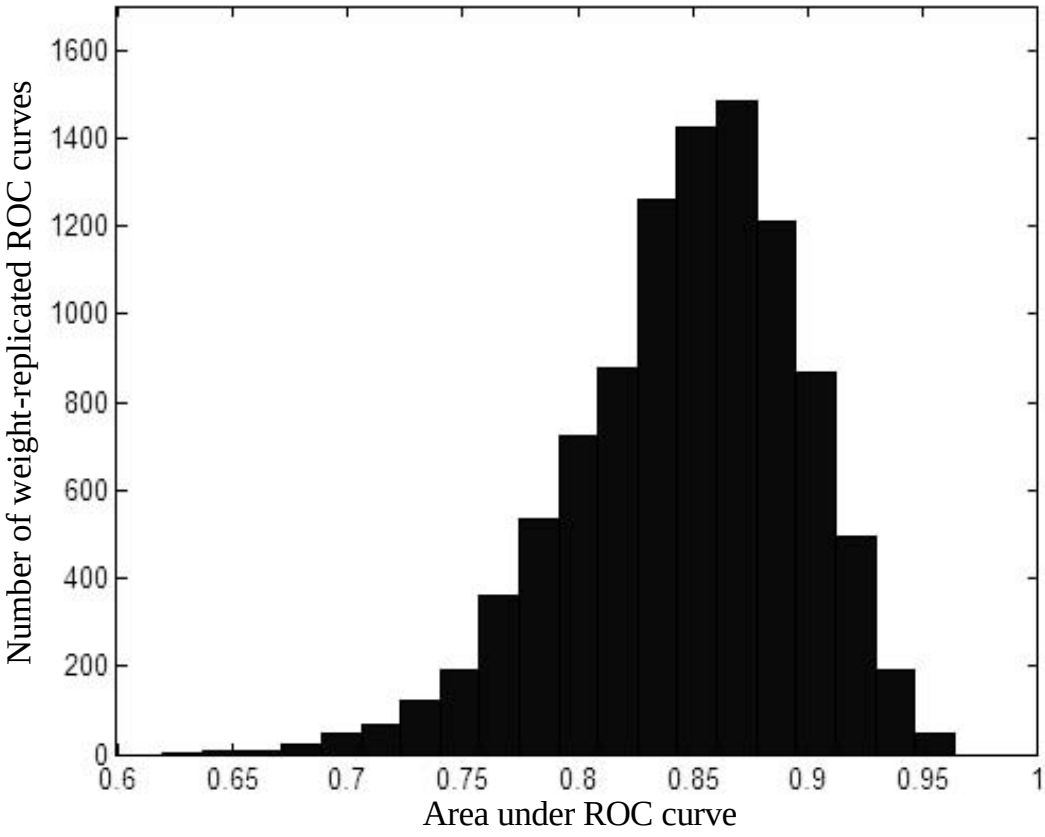


Figure 4.1 An AUC value histogram. This histogram is based on 30 target and 30 non-target samples. After the replication of each representative ROC curve (as in Figure 3.8) a number of times proportional to its weight, an AUC value is calculated for each curve. For this example the underlying densities are known (but not used in the histogram development), and the true AUC value is 0.882. The AUC value is a single summary metric used to compare different SUTs, and here an extention is made to a density estimate in the form of a histogram of AUC values.

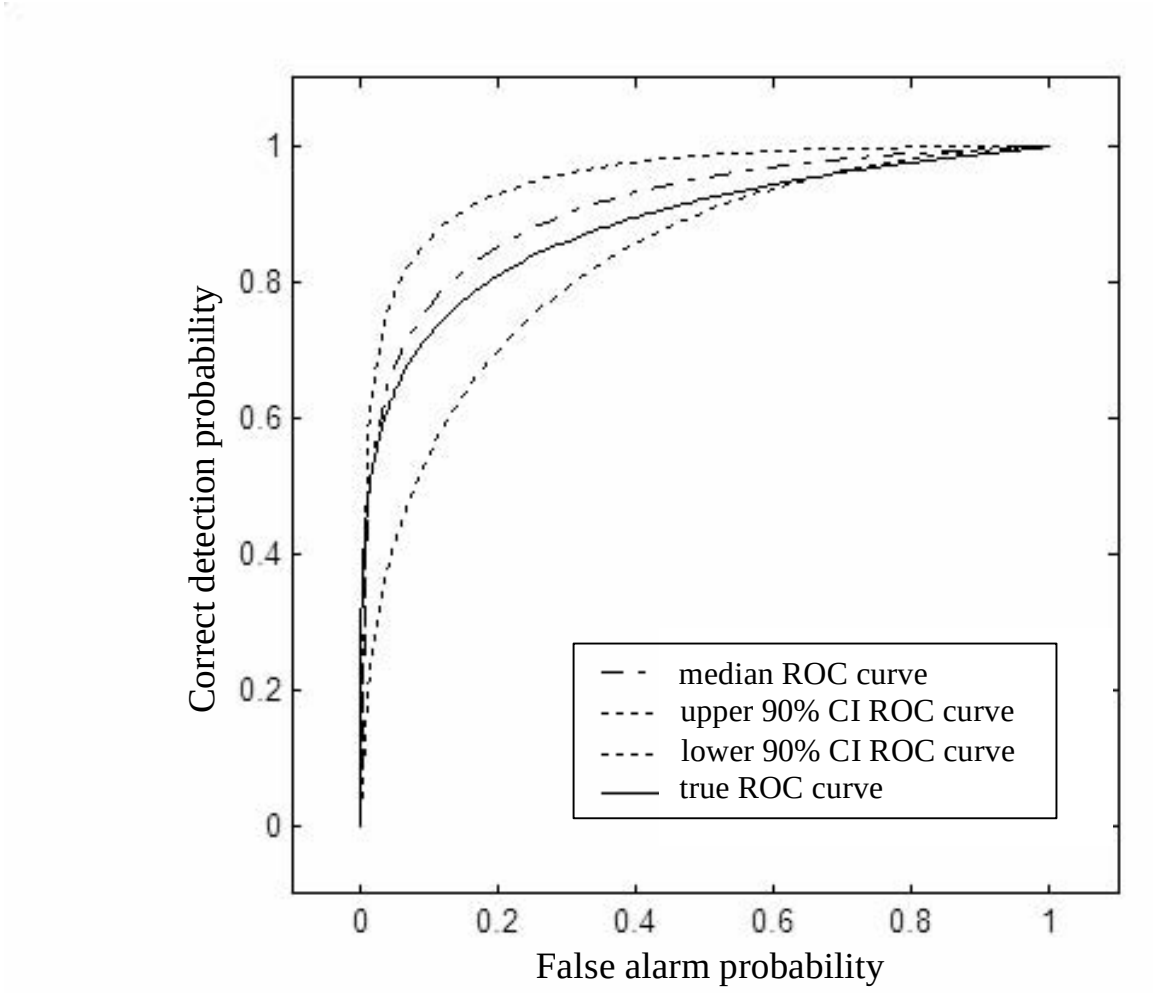


Figure 4.2 Rank characterization for ROC curves weighted by AUC values. Once the ROC curve density is developed, there are many possible definitions of ROC curve confidence bands or ROC curve confidence interval contours. The 90% ROC curve confidence interval contours shown here are obtained by finding the ROC curve that has the AUC value closest to the lower 90% AUC value confidence bound and the ROC curve that has the AUC value closest to the upper 90% AUC value confidence bound. The median ROC curve is the ROC curve that has the AUC value closest to the median (50%) AUC value, and the true ROC curve (the ROC curve for the target and non-target densities from which the samples are drawn) is also shown.

alarm probability, the 90% confidence bands of correct detection probability for two SUTs can be compared. In particular, if one SUT has a median ROC curve which has greater correct detection probability at the selected false alarm probability than a second SUT, and if the confidence intervals of both SUTs at this false alarm probability do not overlap, then the first SUT is more desirable than the second with at least 90% confidence. The confidence interval at false alarm probabilities approaching zero or one necessarily becomes narrow, because a ROC curve by definition has correct detection probability of zero at false alarm probability of zero and correct detection probability of one at false alarm probability of one. In particular, in Equations (2.2) and (2.3) for correct detection and false alarm probability, respectively, let t = -∞ (or in the case of s$\in [0,1]$, let t = 0). Then correct detection probability equals one and false alarm probability equals one. Let t = ∞ (or in the case of s$\in [0,1]$, let t = 1). Then correct detection probability equals zero and false alarm probability equals zero.

The confidence band method that Figure 4.2 illustrates compares favorably with a confidence band formed by a pair of error bar contours, where such contours are based on the standard deviation of the ROC curve density at a given false alarm probability. Such error bars may extend outside the zero to one range of correct detection probability and do not make appropriate allowances for skewed distributions. Methods in the recent literature that go beyond simple error bars (such as [Zhou and Qin, 2005]) may also extend beyond allowed regions, e.g., to correct detection probabilities greater than one. Two advantages of the ROC curve confidence bands described in this section are that they do not require the selection of an independent variable (such as false alarm probability), and the confidence bands generated are true ROC curves.

Once a density of ROC curves is developed, there are many possible definitions of ROC curve confidence intervals or confidence interval bands (in addition to many ways to compute these intervals or bands). Methods described in the literature typically are applicable to only one or a small subset of these definitions. In contrast, the approach

taken here of forming ROC curve densities first and then transitioning to descriptive statistics can handle a variety of definitions. Ma [Ma and Hall, 1993] emphasizes the need for approaches that may be applied to multiple confidence definitions. The next section details the primary method of confidence interval estimation used in this research.

4.1.3 Characterization of ROC curve density. With false alarm probability as the independent variable, the following procedure generates a ROC curve density characterization. First, find the density of correct detection probability at a selected false alarm probability. Second, repeat for all possible false alarm probabilities. Finally, generate a normalized combination of all such densities to form a ROC probability density.

The density of correct detection probability at a given false alarm probability is found as discussed in Section 3.2, where each ROC curve is replicated a number of times proportional to the posterior parameter weighting $w_k w_m$, given by Equation (3.15), and let N_{wroc} equal the number of replicated ROC curves. Note that each ROC curve gives one correct detection probability value at any selected false alarm probability. A density of correct detection probability may be generated by using each of the N_{wroc} correct detection probabilities as observations of some unknown density, where N_{wroc} is the number of replicated ROC curves, and by estimating the density of correct detection probability based on these observations. The upper plot of Figure 4.3 shows such an estimate based on a beta density model, and the lower plot shows contours of equal density. Figure 4.4 shows similar plots for the true ROC curve with a lower AUC value.

The ROC curve density developed here specifies false alarm probability as the independent variable. However, it is also acceptable (although not as consistent with common practice) to select correct detection probability as the independent axis and to find the density of false alarm probability at every correct detection probability.

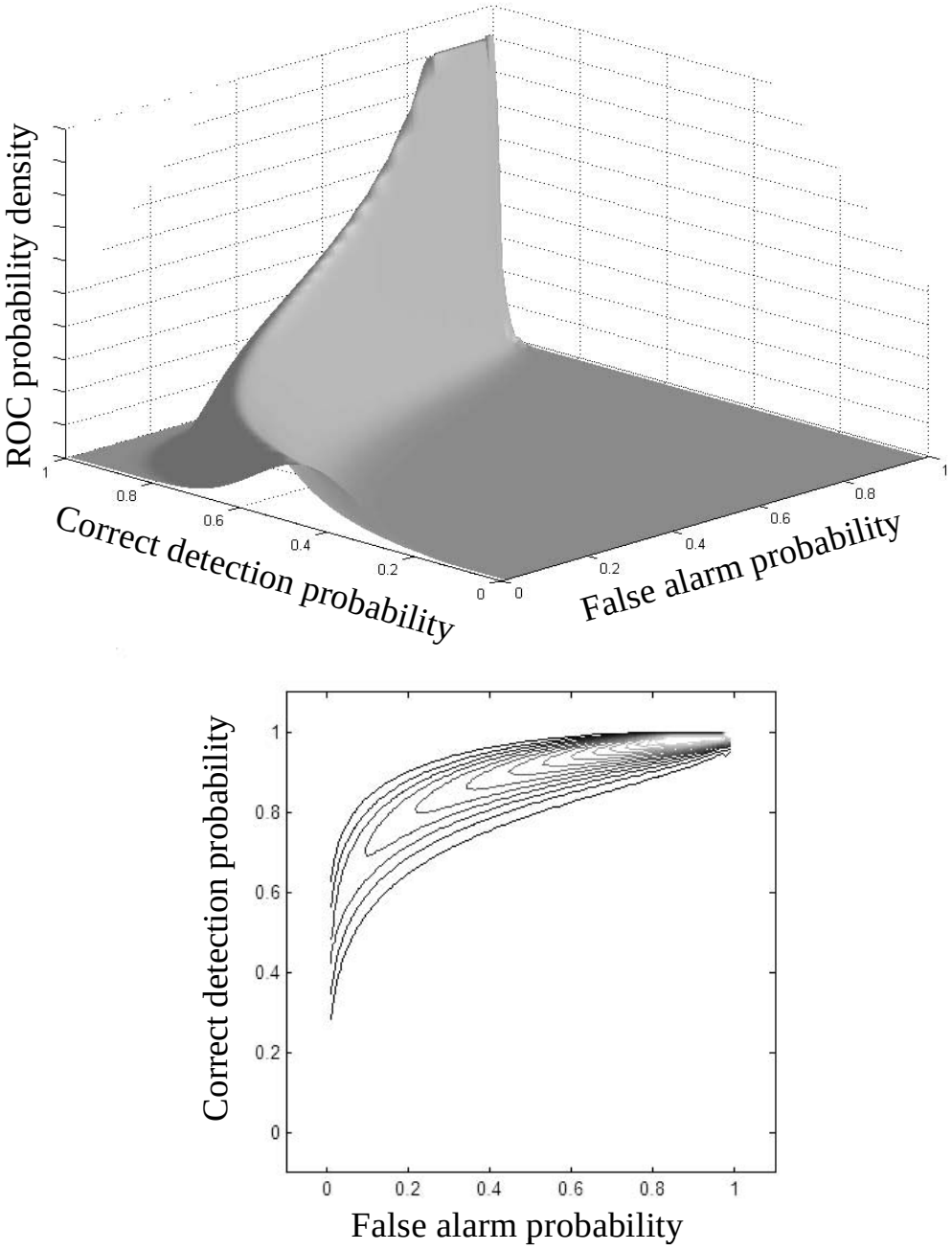


Figure 4.3 A ROC curve density. The upper plot estimates the ROC curve density formed from 30 target scores and 30 non-target scores. Correct detection probability is normalized so that for each false alarm probability the integral of correct detection probability is one. The resulting correct detection density at each selected false alarm probability is smoothed by a beta density that has the same mean and variance as the correct detection probabilities of the replicated ROC curves. The lower plot shows equal density contours.

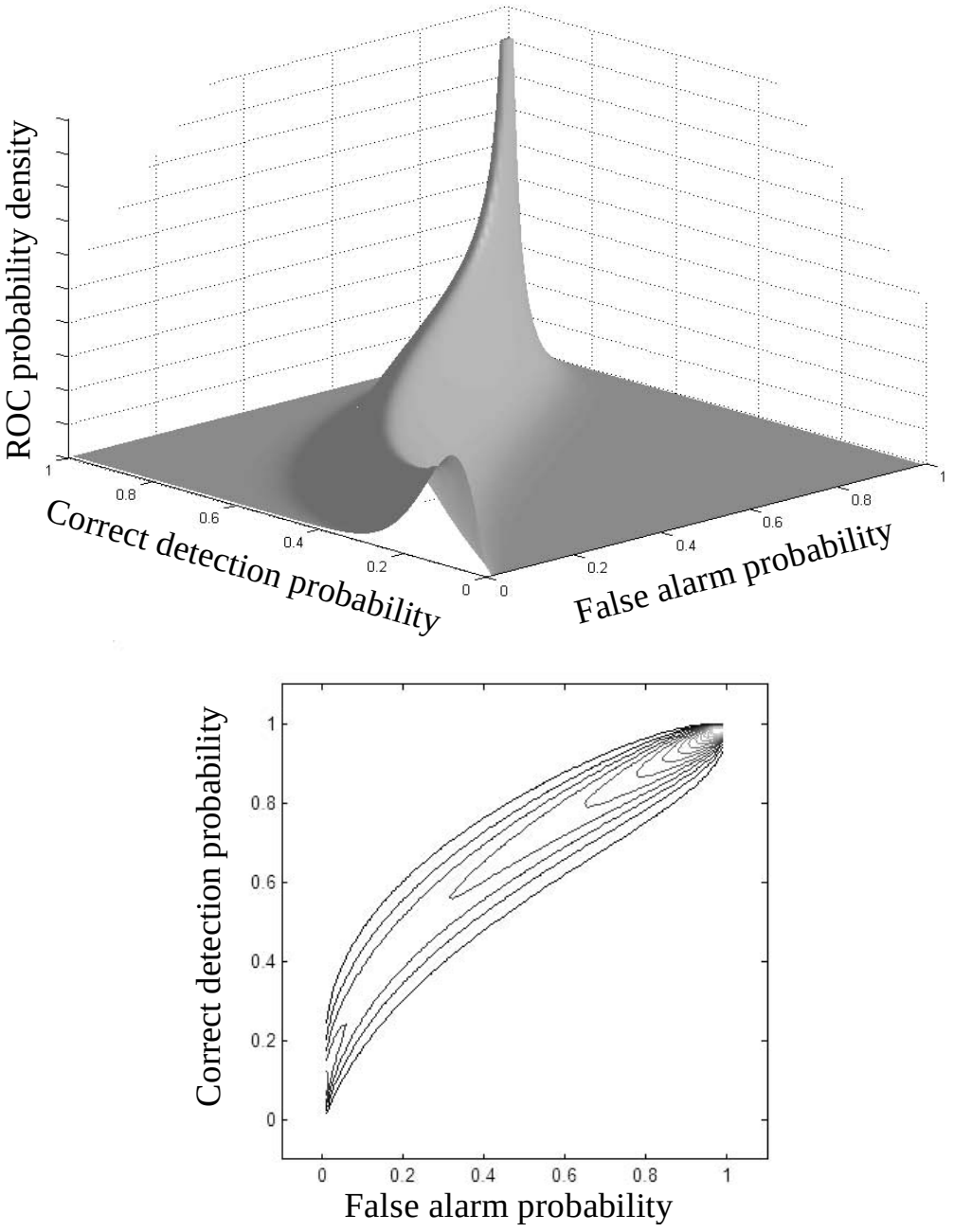


Figure 4.4 A ROC curve density. This figure is similar to Figure 4.3, except that here the set of 30 target scores and 30 non-target scores are selected from different underlying target and non-target densities. These densities are such that the true ROC curve has a lower AUC value than is the case in Figure 4.3.

4.1.4 Confidence contours for ROC curve density. The ROC curve density developed in Chapter 2 permits computation of confidence contours. Consider the N_{wroc} correct detection probabilities at a specified false alarm probability. Create a density based on these N_{wroc} values by centering an impulse (or delta function) density at each of the correct detection probabilities, and normalize the combination of all N_{wroc} impulses so that they form a probability density. Start at a correct detection probability of zero, and increase it until 5% of the correct detection density is enclosed. The correct detection probability where this result occurs is a 90% lower confidence interval. Similarly, start at correct detection probability of one and decrease it until 5% of the correct detection density is enclosed to find a 90% upper confidence interval. Repeat for all false alarm probabilities. The continuum loci of all 90% lower confidence intervals specifies a 90% lower confidence contour, and the loci of all 90% upper confidence intervals specifies a 90% upper confidence contour. The two contours enclose a 90% confidence band, and are shown in the upper and lower plots of Figure 4.5. The upper plot uses 10 target samples and 10 non-target samples as inputs, and the lower plot uses 30 target samples and 30 non-target samples as inputs (these samples are similar to those shown in Figure 4.3).

The contours are expressed as follows. Let $p_{y|x}(y|x,d,h)$ denote the ROC density. Then 90% confidence interval for y at a particular x, or (x_i), for a set of target samples (d) and non-target samples (h) are found using

$$CI_{lower}(m_{lower}; x_i, d, h) = \int_0^{m_{lower}} p_{y|(x,d,h)}(y|x_i,d,h)dy \tag{4.3}$$

$$CI_{upper}(m_{upper}; x_i, d, h) = \int_{m_{upper}}^1 p_{y|(x,d,h)}(y|x_i,d,h)dy \tag{4.4}$$

and solving for $CI^{-1}_{lower}(0.05; x_i, d, h)$ and $CI^{-1}_{upper}(0.05; x_i, d, h)$.

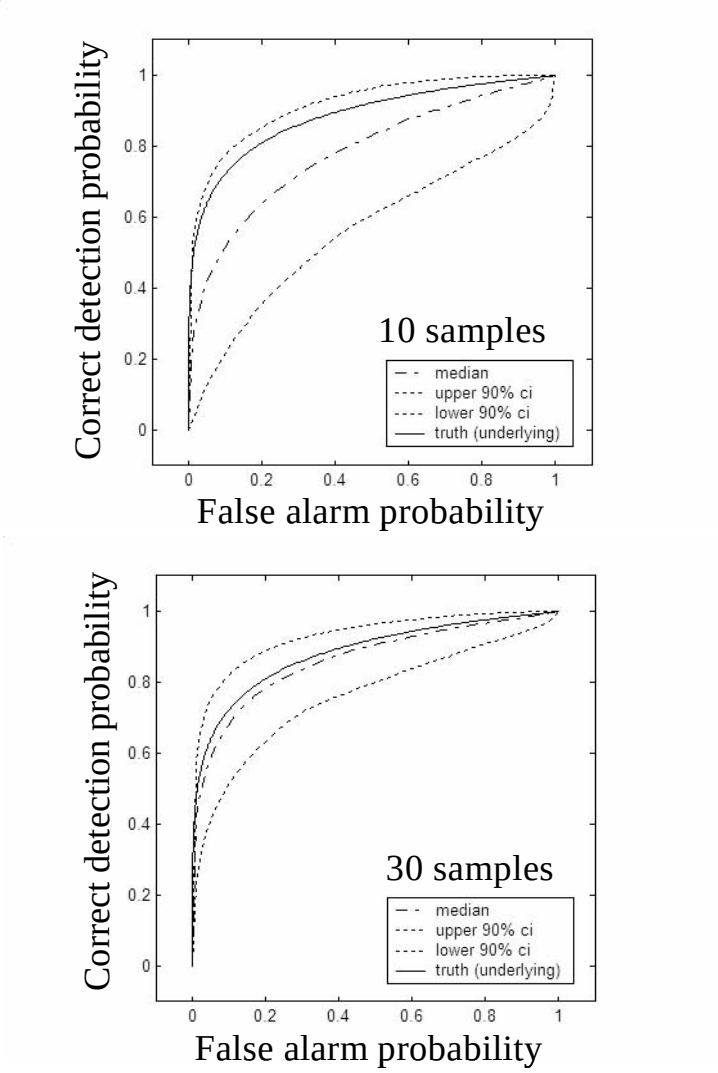


Figure 4.5 Confidence intervals with false alarm probability as the independent variable for two sample sizes. A 90% lower confidence interval is developed from the ROC curve density by fixing a false alarm probability, starting at a correct detection probability of zero, and increasing the correct detection probability until 5% of the density area is encompassed. Similarly, a 90% upper confidence interval is developed by fixing a false alarm probability, starting at a correct detection probability of one, and decreasing the correct detection probability until 5% of the total correct detection probability is encompassed. The median contour (i.e., the locus of points that encompass 50% of correct detection probability) and the true ROC curve (for the target and non-target densities from which the samples are drawn) are also shown. In the upper plot 10 samples of target and 10 samples of non-target are used, and in the lower plot 30 samples of target and 30 samples of non-target are used.

Figure 4.5 shows the general effect on confidence interval of an increase in sample size. The confidence interval widths become smaller as the number of samples increases. The factors that regulate ROC curve density (and confidence interval) widths are $\prod_i f(s_i;u) \prod_j g(s_j;v)$,where f is the non-target density, g is the target density, s_i are the non-target samples, s_j are the target samples, and u and v are the specified parameters that define f and g. As the number of target and non-target samples increases, the range of densities with a high weight value as defined by these function decreases.

The density-based ROC curve confidence interval generation method developed here constitutes an improvement over other methods described in the literature, as the intervals here have more useful definitions. Many existing methods attempt to describe the uncertainty in probability of correct detection y at a specific probability of false alarm x, but do not permit extrapolation to confidence bands because they either fail to incorporate or incorporate conservatively the underlying uncertainty in the variable x. The non-target density yields this uncertainty as a simple outcome of the Bayesian approach in the method developed here. Other existing methods incorporate uncertainty in both y and x, but restrict threshold to a single value or make assumptions that are only valid for particular density forms (see [Linnet, 1987], [Campbell, 1994], and [Platt *et al.*, 2000]). In the method described here, threshold is eliminated as a variable, which removes the need to restrict threshold to a single value and retains uncertainty in the independent variable x.

A confidence accuracy measure designated alpha tests ROC curve confidence interval accuracy. Alpha describes the percentage of trials where the confidence interval does not enclose truth. One set of target samples and non-target samples define one trial, a second set of target samples and non-target samples define a second trial, etc. An ideal alpha is one minus the intended confidence interval coverage. The example in Figure 4.5 claims 90% confidence intervals, and thus the ideal alpha is 0.1. If the underlying target and non-target densities generate the same number of target and non-target samples an

infinite set of times, the truth ideally departs from the confidence interval 10% of the time. This confidence region accuracy evaluation process extends to a confidence band, where contours defined by the confidence intervals at every false alarm probability define the band. The process here assumes that each false alarm probability has an equal contribution to an overall alpha measure. For example, if the true ROC curve lies outside the generated confidence band for 25% of the false alarm probabilities for each of an infinite set of ROC curve estimates, then alpha is 0.25. An alternative approach declares "failure" if any portion of the ROC curve confidence band lies outside of the confidence band for any false alarm probability for a particular run. With this alternative approach, if any portion of the true curve deviates from the ROC curve confidence band on 40% of an infinite set of generated ROC curve confidence bands, then alpha is 0.40.

Confidence interval accuracy does not necessarily increase with increase in sample size. Consider two extreme cases. First, evaluate a ROC curve estimate with infinitely small confidence interval widths that are ideally 90% confidence intervals. The ROC curve estimate may be close to truth, but the confidence band is always above or below the true ROC curve, resulting in an average alpha of 1. Next, consider a ROC curve estimate far from truth, but which has the largest possible confidence interval widths. For example, at every false alarm probability, the 90% confidence interval limits are 0 and 1, in which case alpha is 0. In a related consideration, note that a confidence interval calculation approach that produces an alpha of 0.1 (for claimed 90% confidence intervals) is generally better than an approach that produces an alpha of 0.

Let $r_{true}(x)$ be the true ROC curve (in test cases where the density that generates the target and non-target samples is known), let $ca(x)$ be the actual coverage accuracy defined by Equation (4.5), let $CI_{lower}(m;x)$ and $CI_{upper}(m;x)$ be as defined by Equations (4.3) and (4.4). Then

$$ca(m,x) = P\left\{CI_{lower}(m;x) < r_{true}(x) < CI_{upper}(m;x)\right\}. \tag{4.5}$$

Estimates for $P(r_{true}(x) > CI_{lower}(m;x))$ and $P(r_{true}(x) < CI_{upper}(m;x))$ may be found by generating many sets of identical numbers of samples from the same target and non-target score densities, to approximate the probabilities noted in Equation (4.5). In particular, let $\text{ccv}_{desired}$ be the desired confidence interval coverage (in the case of 90% confidence intervals, $\text{ccv}_{desired} = 0.90)$, and let ad, $alpha\ desired$, be one minus the desired confidence interval coverage. Then

$$alpha(m) = 1 - \int_0^1 [ca(m,x) - ad]dx. \tag{4.6}$$

4.1.5 Relations of confidence intervals to Chebyshev's inequality. Three separate relations of Chebyshev's inequality to confidence intervals follow.

The first relation is established in Theorem 4.1 and shows that the upper and lower bounds of the confidence interval contours developed in Section 4.1.4 are within the constraints established by Chebyshev's inequality.

Theorem 4.1 Upper and lower bounds for confidence interval contours

Let $p_{y|x}(y|x)$ be as developed in Theorem 3.2. The median (see [DeGroot and Schervish, 2002, pp. 210]) of $p_{y|x}(y|x)$ is the value $med_{y|x}$ such that

$$\int_0^{med_{y|x}} p_{y|x}(\eta|x)d\eta = \int_{med_{y|x_i}}^1 p_{y|x}(\eta|x)d\eta = 0.5. \tag{4.7}$$

Let $p_{up_{y|x}}(y|x)$ and $p_{low_{y|x}}(y|x)$ be symmetric probability densities such that

$$p_{up_{y|x}}(y|x) = \left\{ \begin{array}{c} p_{y|x}(y|x) \forall y \geq med_{y|x} \\ p_{y|x}((2med_{y|x} - y)|x) \forall y < med_{y|x} \end{array} \right\} \tag{4.8}$$

and

$$p_{low_{y|x}}(y|x) = \left\{ \begin{array}{c} p_{y|x}(y|x) \forall y \leq med_{y|x} \\ p_{y|x}((2med_{y|x} - y)|x) \forall y > med_{y|x} \end{array} \right\}. \tag{4.9}$$

Also, let $\mu_{p_{up_{y|x}}}(y|x) =$ mean of $p_{up_{y|x}}(y|x)$, $\mu_{p_{low_{y|x}}}(y|x) =$ mean of $p_{low_{y|x}}(y|x)$, $\sigma_{p_{up_{y|x}}}(y|x)$ = standard deviation of $p_{up_{y|x}}(y|x)$, and $\sigma_{p_{low_{y|x}}}(y|x)$ = standard deviation of $p_{low_{y|x}}(y|x)$. Finally let $r_u(x)$ denote the upper bound on the (1 - $alpha$) upper confidence interval of $p_{y|x}(y|x)$ and let $r_l(x)$ denote the lower bound on the (1 - $alpha$) upper confidence interval of $p_{y|x}(y|x)$.

Then

$$r_u(x) \leq med_{y|x} + (\sigma_{p_{up_{y|x}}}(y|x) \sqrt{\frac{2}{alpha}}), \tag{4.10}$$

$$r_l(x) \geq med_{y|x} - (\sigma_{p_{low_{y|x}}}(y|x) \sqrt{\frac{2}{alpha}}). \tag{4.11}$$

Proof

By Chebyshev's inequality (see [Hogg and Craig, 1978, pp. 59]), for $k > 0$

$$P(r_u(x) - \mu_{p_{up_{y|x}}}(y|x) \geq k \sigma_{p_{up_{y|x}}}(y|x)) \geq 1 - \frac{1}{k^2}. \tag{4.12}$$

Thus

$$P(r_u(x) \geq k \sigma_{p_{up_{y|x}}}(y|x) + \mu_{p_{up_{y|x}}}(y|x)) \geq 1 - \frac{1}{k^2}. \tag{4.13}$$

An upper bound on the (1-$alpha$) upper confidence interval specifies that
$P(r_u(x) \geq k\sigma_{p_{up_{y|x}}}(y|x) + \mu_{p_{up_{y|x}}}(y|x)) \geq 1 - \frac{alpha}{2}$

and thus $1 - \frac{1}{k^2} = 1 - \frac{alpha}{2}$, $k = \sqrt{\frac{2}{alpha}}$,

Here $p_{up_{y|x}}(y|x)$ is symmetric in y, $\mu_{p_{up_{y|x}}}(y|x) = med_{y|x}$, and by definition, $r_u(x)$ denotes the (1 - $alpha$) upper confidence interval.

Thus $r_u(x) \leq med_{y|x} + (\sigma_{p_{up_{y|x}}}(y|x)\sqrt{\frac{2}{alpha}})$.

Similarly, by Chebyshev's inequality (see [Hogg and Craig, 1978, pp. 59]),

$$P(\mu_{p_{up_{y|x}}}(y|x) - r_l(x) \geq k\sigma_{p_{low_{y|x}}}(y|x)) \geq 1 - \frac{1}{k^2}, \tag{4.14}$$

$$P(-r_l(x) \geq k\sigma_{p_{low_{y|x}}}(y|x) - \mu_{p_{low_{y|x}}}(y|x)) \geq 1 - \frac{1}{k^2}, \tag{4.15}$$

and

$$P(r_l(x) \leq \mu_{p_{low_{y|x}}}(y|x) - k\sigma_{p_{low_{y|x}}}(y|x)) \geq 1 - \frac{1}{k^2}. \tag{4.16}$$

A lower bound on the (1-$alpha$) lower confidence interval specifies that
$P(r_l(x) \leq \mu_{p_{low_{y|x}}}(y|x) - k\sigma_{p_{low_{y|x}}}(y|x)) \geq 1 - \frac{alpha}{2}$.

Here $p_{up_{y|x}}(y|x)$ is symmetric in y, $\mu_{p_{low_{y|x}}}(y|x) = med_{y|x}$, and

by definition, $r_l(x)$ denotes the (1 - $alpha$) upper confidence interval.

Thus, $r_l(x) \leq med_{y|x} + (\sigma_{p_{low_{y|x}}}(y|x)\sqrt{\frac{2}{alpha}})$.

Figure 4.6 shows a plot with the 90% confidence intervals developed in Section 4.1.4 and the upper and lower bounds for the upper and lower 90% confidence intervals as developed in this Section.

The second relation of confidence intervals to Chebyshev's inequality does not require the Bayesian progression that is the focus of the research presented here, but it results in extremely wide (and unformative) confidence bounds. This relation is established as follows.

For a given set of target samples d, a given set of non-target samples h, and a selected $alpha$ (such as $alpha = 0.1$), find a target sample standard deviation $\widehat{\sigma}_t$ and a non-target sample standard deviation $\widehat{\sigma}_n$, and find the upper and lower bounds on the target mean as follows. From Chebyshev's inequality (see [Hogg and Craig, 1978, pp. 59]),

$$P(|\text{mean}(d) - x_t| < k\widehat{\sigma}_t) \geq 1 - \frac{1}{k^2} = (1 - alpha). \tag{4.17}$$

Find the two values of x_t such that

$$|\text{mean}(d) - x_t| < k\widehat{\sigma}_t. \tag{4.18}$$

Similarly, find the upper and lower bounds on the non-target mean by solving for x_n, where

$$P(|\text{mean}(h) - x_n| < k\widehat{\sigma}_n) \geq 1 - \frac{1}{k^2} = (1 - alpha), \tag{4.19}$$

and find the two values of x_n such that

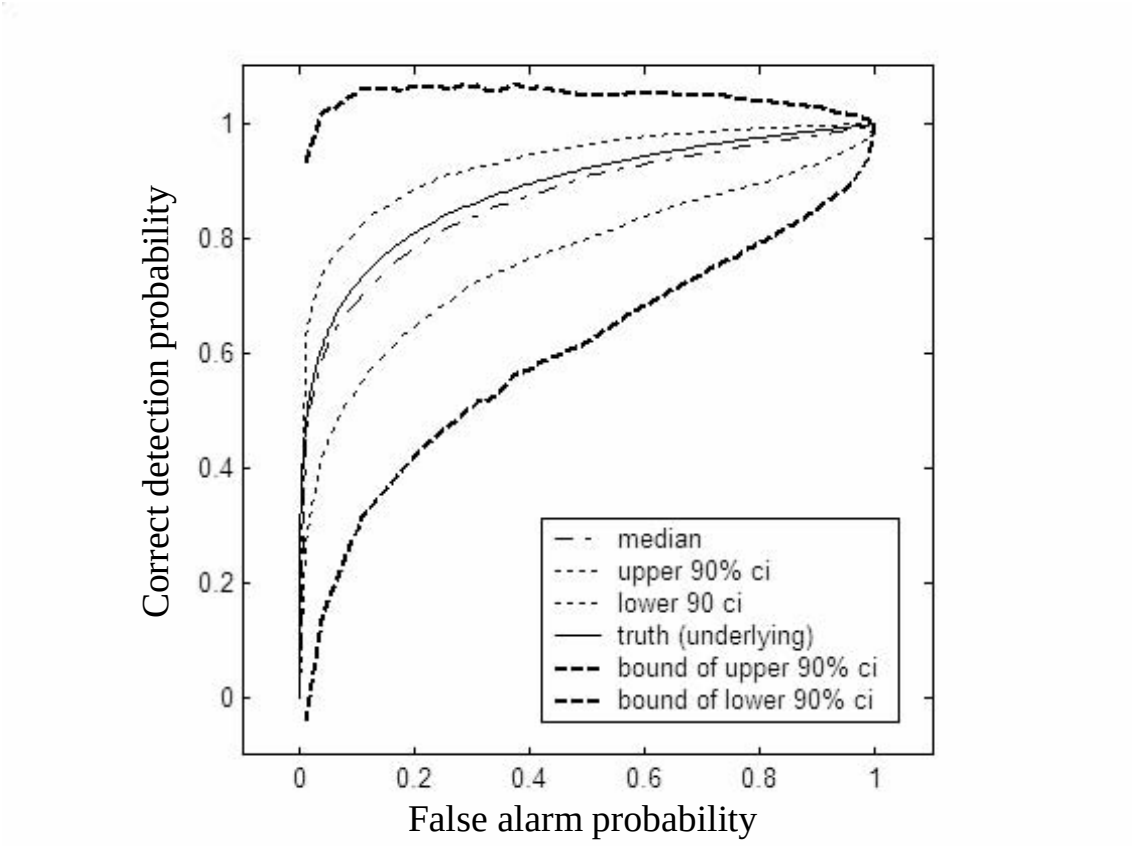


Figure 4.6 Upper and lower bounds on 90% confidence intervals plus ROC curves and coverage for a selected density pair. Here beta target and non-target densities generate 30 target and 30 non-target samples (the densities have $\mu = 0.805$, $\sigma = 0.059$ and $\mu = 0.715, \sigma = 0.046$, respectively). The 90% confidence intervals for the ROC curve developed using the method described in Section 4.1.4 are the short dashed curves. The underlying true ROC curve is the solid curve, the median ROC curve estimate is the dash-dotted curve, and the upper and lower bounds of the 90% confidence intervals are the heavy dashed curves.

$$|\text{mean}(h) - x_n| < k\widehat{\sigma}_n. \tag{4.20}$$

This approach results in ROC curve uncertainty estimates that are extremely wide and uninformative, even when the target and non-target standard deviations are specified. If uncertainty in the target and non-target standard deviations is incorporated, these bounds will only become wider and less informative. Figure 4.6 provides an example for 30 target and 30 non-target samples. Here it is assumed that the standard deviation is constant at the standard deviation of the target and non-target samples, and a target and non-target beta density model is assumed (both of these selections can only narrow the bands compared with more general cases). In combinations where the mean and standard deviation pairs are outside of the admissible set (of allowable means and standard deviations for a beta density), the standard deviation is retained, but the mean is adjusted (brought closer to the sample mean) so that the resulting mean and standard deviation are within the admissible set. This adjustment can only make the calculated bounds more narrow.

Finally, a third relation of confidence intervals to Chebyshev's inequality solves for m_{lower} and m_{upper} such that (for 90% confidence bounds)

$$CI_{lower}(m_{lower}; x, d, h) = \int_0^{m_{lower}} p_{y|(x,d,h)}(y|x,d,h)dy = 0.05 \tag{4.21}$$

$$\text{and } CI_{upper}(m_{upper}; x, d, h) = \int_{m_{upper}}^1 p_{y|(x,d,h)}(y|x,d,h)dy = 0.05, \tag{4.22}$$

where m_{lower} is the correct detection probability y that produces a 5% lower confidence interval at a specified false alarm probability x, for a set of target samples h and a set of non-target samples d, and m_{upper} is the correct detection probability y that produces a 5%

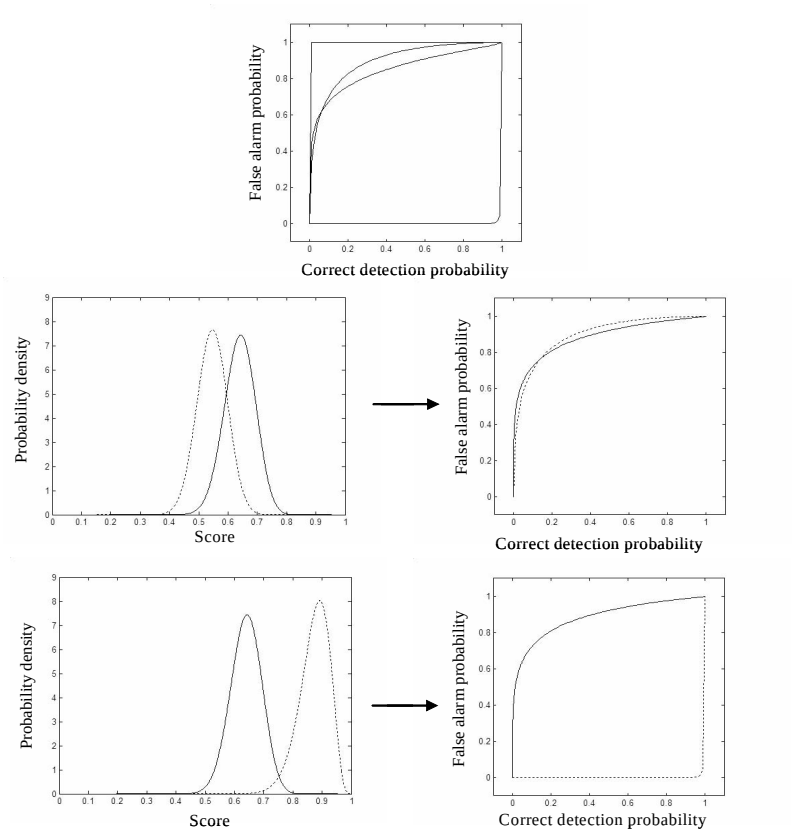


Figure 4.7 ROC curve uncertainty example with Chebyshev's inequality. ROC curve estimates are produced from the underlying target and non-target densities of Figure 4.6. Equations 4.17 through 4.20 are applied to find the 90% bounds on uncertainty of the target and non-target means. The standard deviation of the target and non-target samples is used, and target and non-target densities at the extremes of the uncertainty bounds are combined to form the curves shown in the top plot. The upper and lower limits of these curves form confidence bounds; these bounds are extremely wide (the upper ROC curve has an AUC value ≈ 1, and the lower ROC curve has an AUC value ≈ 0). The four lower plots show two of the four sets of density pairs at the uncertainty bound extremes. In the bottom right plots, the ROC curves that correspond with the underlying target and non-target densities are shown as solid curves, and the curves that correspond with the densities at left are shown as dotted curves.

upper confidence interval at a specified false alarm probability x, for a set of target samples h and a set of non-target samples d.

From Chebyshev's inequality (see [Hogg and Craig, 1978, pp. 58])

$$P[m_{lower}(x;d,h) \leq c_{lower}(x)] \geq \frac{E[m_{lower}(x;d,h)]}{c_{lower}(x)}, \tag{4.23}$$

$$\text{and } P[m_{upper}(x;d,h) \geq c_{upper}(x)] \leq \frac{E[m_{upper}(x;d,h)]}{c_{upper}(x)}, \tag{4.24}$$

where $c_{lower}(x)$ and $c_{upper}(x)$ are lower limits to the lower 90% confidence interval and upper limits to the upper 90% confidence interval. This progression requires the calcluation, based on one set of target and non-target samples, of the expected value of $m_{lower}(x;d,h)$ and $m_{upper}(x;d,h)$. Based on one set of target and non-target samples, the best estimate is $E[m_{lower}(x;d,h)] = m_{lower}(x;d,h)$, and $E[m_{upper}(x;d,h)] = m_{upper}(x;d,h)$. If more sets of samples are available, then these new samples may be incorporated into the framework, and improved confidence intervals may be developed. However, $p_{y|(x,d,h)}(y|x,d,h)$ is already the defined (actual) posterior probability density for the ROC curve that fully incorporates what is known from the observed target and non-target samples (which are assumed independent and identically distributed), assumed model, and assumed priors. Thus, this discussion indicates that the target and non-target samples d and h are realizations of random variables, and as such the developed posterior probability density, $p_{y|(x,d,h)}(y|x,d,h)$ may be (and should be) updated if additional sets of representative target and non-target samples are available. In any case, the developed posterior probability densities (and the corresponding confidence intervals $CI_{lower}(m_{lower};x,d,h)$ and $CI_{upper}(m_{upper};x,d,h)$) are actual confidence intervals based on the available samples, assumed model, and assumed priors.

The above discussion indicates that the posterior probability density is a full summary of what is known about the ROC curve based on the observed sample data, the assumed model, and the assumed priors. Carlin writes [Carlin and Louis, 2000, pp. 36] that a Bayesian approach "enables direct probability statements about the likelihood of θ falling in [set] C, i.e., 'The probability that θ lies in [set] C given the observed data y is at least (1-α).' This is in stark contrast to the usual frequentist CI, for which the corresponding statement would be something like, If we could recompute [set] C for a large number of datasets collected in the same way as ours, about (1-α) x 100% of them would contain the true value of θ.' This is not a very comforting statement, since we may not be able to even imagine repeating our experiment a large number of times" (the use of [set], in brackets, has been inserted here for clarity). This discussion by Carlin is applicable to the research presented here if the actual ROC curve is denoted as θ, if C is the set of all real values such that $m_{lower} \leq C \leq m_{upper,}$ if y refers to the observed target and non-target samples, and if $\alpha = 0.1$ (for 90% confidence intervals). MacKay [MacKay, 2003, pp. 50] summarizes the value of the posterior probability distribution strongly in the following statement: "The posterior probability distribution represents the unique and complete solution to the problem. There is no need to invent 'estimators'; nor do we need to invent criteria for comparing alternative estimators with each other."

4.1.6 Convergence as number of parameter points increases. Wide spacing between the prior beta density mean and standard deviation points for target densities and/or non-target densities can affect the size of the confidence band. As this spacing approaches zero and as the number of points selected therefore approaches infinity, the confidence band area converges to a constant (the convergence of ROC curve density is proven in Chapter 3; the confidence intervals are then deterministic from this density). A simple example of this process is shown in Figure 4.8. Both plots have as inputs the same 30 target samples and the same 30 non-target samples. The plot at the top, labeled coarse spacing, develops confidence interval contours using the nine highest-weighted points

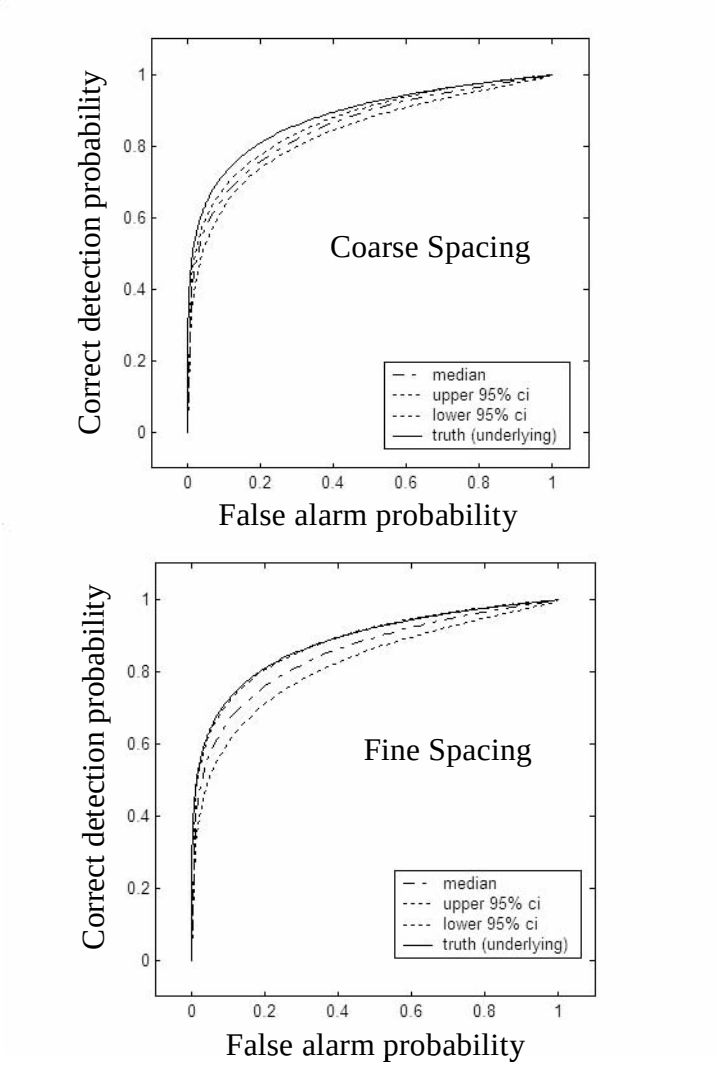


Figure 4.8 The ROC curve confidence interval bands versus spacing of prior beta density mean and standard deviation values. As spacing decreases and the corresponding number of mean and standard deviation values considered therefore increases, the confidence band area converges to a limit. An example of this trend for a 95% confidence interval with false alarm probability as the independent variable is shown here. Both plots use as inputs the same 30 target samples and the same 30 non-target samples. The plot labeled coarse spacing develops confidence intervals using the nine highest-weighted points uniformly spaced on the mean and standard deviation target and non-target beta density axes such that the ratio of the weight of the lowest to the highest points is 0.001. The plot labeled fine spacing develops confidence intervals using the 25 highest-weighted points uniformly spaced on the mean and standard deviation target and non-target beta density axes such that the ratio of the weight of the lowest to the highest points is 0.001.

uniformly spaced in target density mean and standard deviation such that the ratio of the posterior density (or weight) of the lowest to the highest is 0.001. These contours define a confidence band. Nine highest-weighted points are similarly found for the non-target density. Note that if only one point for target density and one point for non-target density is used, the confidence band area is 0 because the ROC curve is deterministic. The plot at bottom, labeled fine spacing, develops a similar confidence band.

Figure 4.9 shows confidence band area convergence as the number of evaluated points increases. For the example in Figure 4.9, target standard deviation versus mean grid points are selected, where these points are centered around the mean and standard deviation of the target samples. The number of target parameter density points is increased from 9 points (3 target means and 3 target standard deviations) to 25 points (5 target means and 5 target standard deviations), etc., up to a total of 1089 points (33 target means and 33 target standard deviations). Each set of points is used to calculate confidence bands. The confidence band area converges (the convergence of ROC curve density is proven in Section 3.2; the confidence intervals are then deterministic from the density) as the number of parameter points increases, which indicates that point spacing does not bias the prior parameter densities if the points are selected uniformly over the target and non-target density parameters (such as mean and standard deviation).

4.1.7 Additional confidence bound definitions. Note that the method developed here extends to an additional class of confidence bounds that are not described elsewhere. These confidence bounds describe ROC curves for a threshold selected at random, with uniform probability of selection over allowable thresholds, where the bounds are formed such that the integral of the ROC curve density above a specified value has the given percentage of unit density variance. Such bounds are an extension of the method developed here.

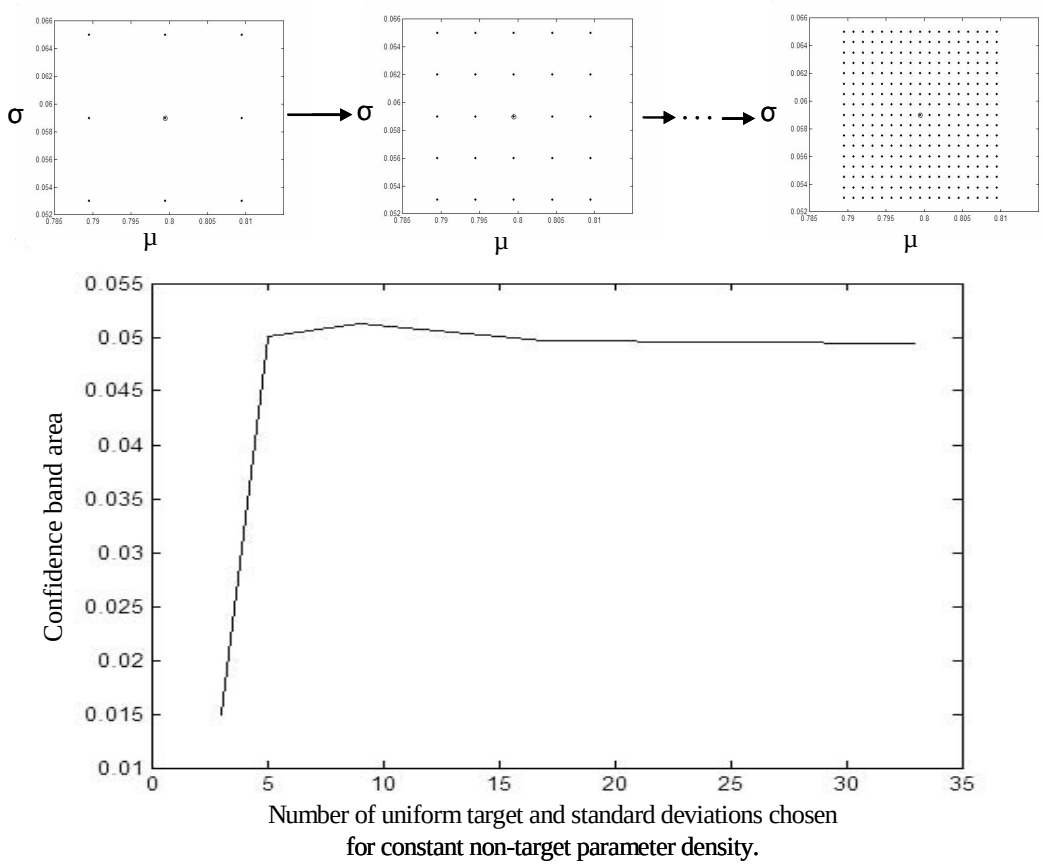


Figure 4.9 Confidence band area versus number of evaluated points. Here a beta score target density with mean of 0.805 and standard deviation of 0.059 and a beta non-target score density with mean of 0.715 and standard deviation of 0.046 generate 300 target and 300 non-target samples. The method of Section 4.1.3 estimates the ROC curve confidence band. The non-target posterior parameter density is evaluated at a single point. The target density is modeled by 3 means and 3 standard deviations (9 points), 5 means and 5 standard deviations, etc., where the mean and standard deviations of the selected points for the 3 mean and 3 standard deviation, 5 mean and 5 standard deviation, and 33 mean and 33 standard deviation cases are shown in the upper two plots. As the number of target parameter points increases, the lower plot shows that the confidence band area approaches a constant.

Let $p_{y,x}(y,x)$ be the joint density of the ROC curve (rather than the ROC curve density normalized such that the probability density of correct detection at given false alarm probability is one), as described by Equation (3.17), except here replace $p_{y|x}(y|x)$ with $p_{y,x}(y,x)$. Let $c.c.$ be the desired coverage (e.g., 0.90). Let the ROC subset ($\mathfrak{S}(z_1)$) be the subset of all x, y pairs such that

$$\mathfrak{S}(z_1) = \left\{ (x,y) \in [0,1]^2 : p_{y,x}(y,x) \geq z_1 \max_{x,y}[p_{y,x}(y,x)] \right\} \tag{4.25}$$

where $x \in [0,1]$, $y \in [0,1]$, and $z_1 \in [0,1]$. Let z_1 = 1 and find

$$c.c.test = \iint_{\mathfrak{S}(z_1)} p_{y,x}(y,x)dxdy. \tag{4.26}$$

Then let z_{1new}=$z_{1old} - \varepsilon$ if $c.c.test < c.c.$. Re-define the ROC subset $\mathfrak{S}(z_{1new})$ for this z_{1new}. Repeat the process, continuing to reduce z_1 until $c.c.test = c.c.$. The subset of all x, y pairs that make up $\mathfrak{S}(z_1)$ where $c.c.test = c.c$ forms the *confidence bound.*

Figure 4.10 shows ROC confidence bounds based on this definition and indicates higher densities close to the ROC extremes. This result is appropriate because any ROC curve has a correct detection probability of zero at false alarm probability of zero and a correct detection probability of one at false alarm probability of one.

4.2 Verification of results

4.2.1 Analysis of ROC curve and AUC value bias. The results that follow quantify the confidence band accuracy for the method described here (in Section 4.1.3) by considering repeated runs over many sets of samples. Before examining this accuracy, consider that ROC curves and AUC values formed by fitting beta densities to beta density generated score samples generally have low bias, even for low numbers of samples. For example (see Figure 4.11), select a target and non-target beta density pair. Generate 30 target and

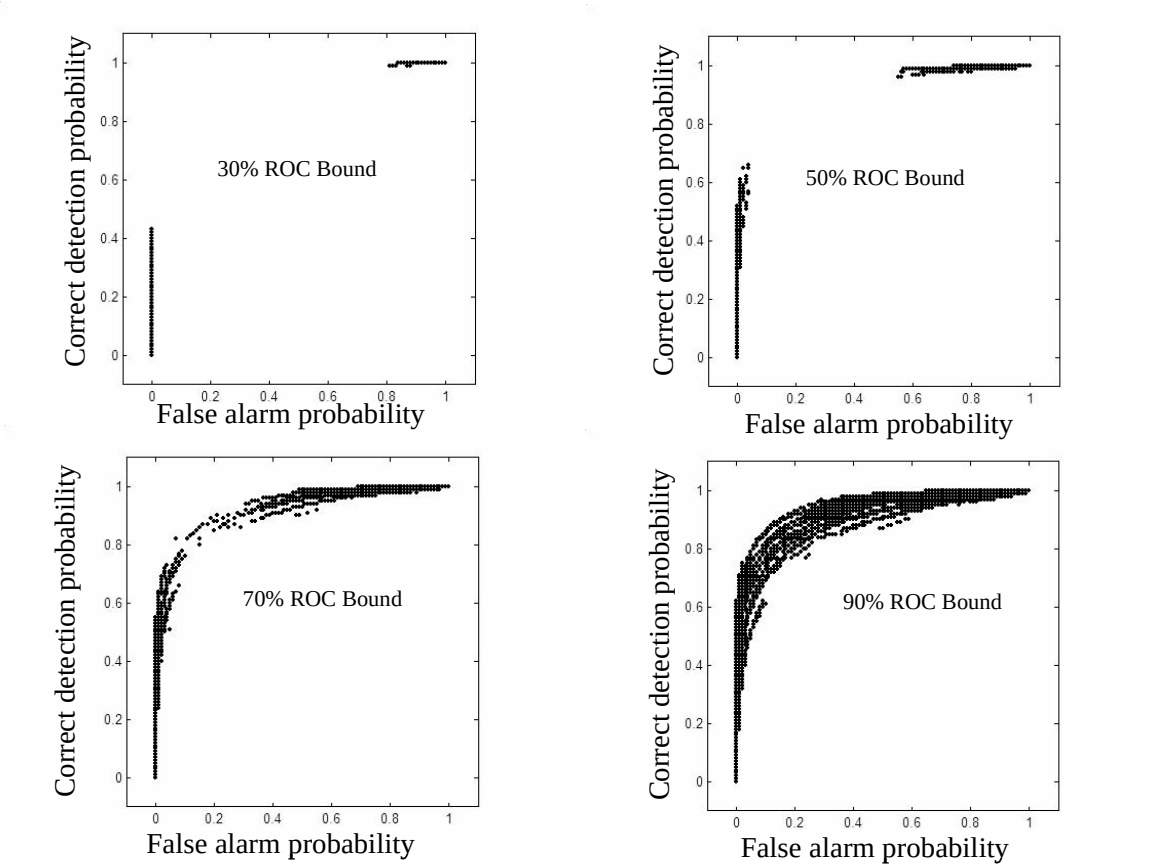


Figure 4.10 The ROC curve uniform threshold confidence bounds. The four plots show 30%, 50%, 70%, and 90% ROC curve bands formed such that the integral of the ROC curve density above a specified value has the given percentage of unit density volume, assuming that score threshold is randomly and uniformly selected over all allowed threshold values (0 to 1). Note that only the 2-D area of showing the region bounded by this 3-D density is shown in the above plot.

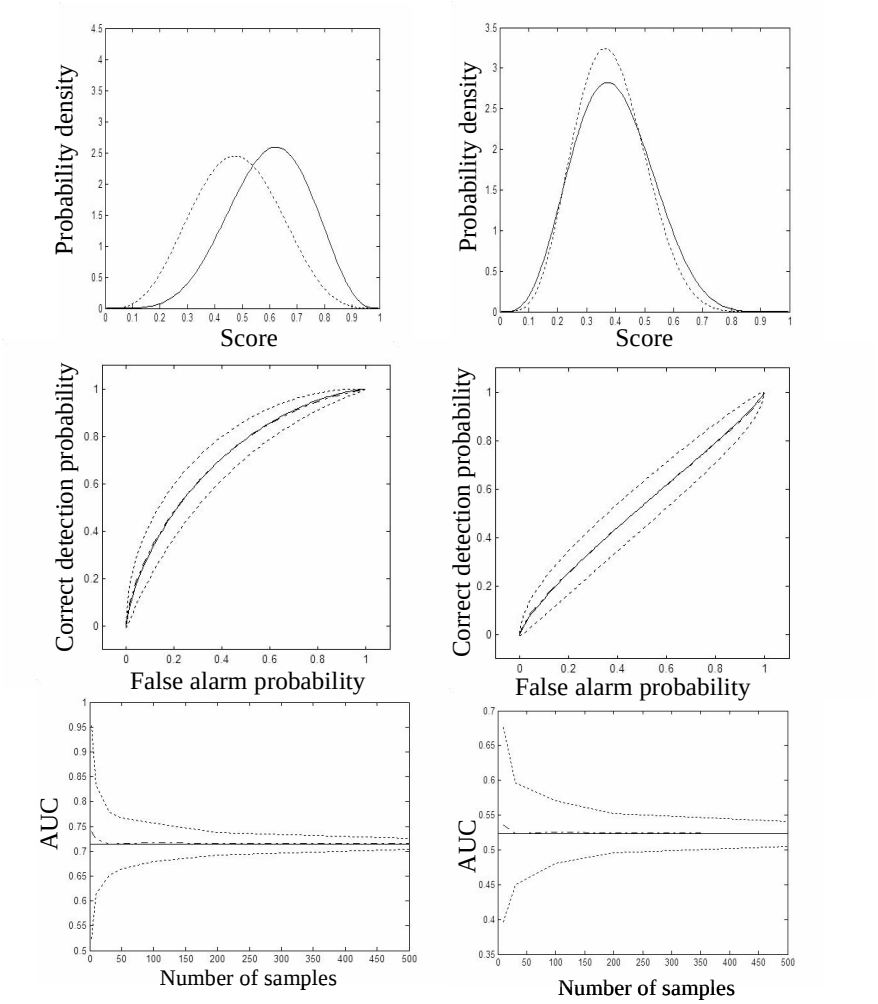


Figure 4.11 Estimates of ROC curves and AUC values from mean and variance of target and non-target beta densities. The top two plots show the underlying beta target densities (solid curves) and the underlying beta non-target densities (dashed curves); the respective mean and standard deviation parameters are 0.599, 0.021, and 0.479, 0.023. The middle left plot shows the ROC curve for the underlying beta densities (solid curve) with ROC curve statistics for 300 sets of 30 target and 30 non-target samples drawn from each density, where the mean of the 300 curves (dash/dotted line) and this mean plus and minus the standard deviations are plotted (dotted lines). The lower left plot similarly shows the true AUC value, mean AUC value, and mean AUC value plus and minus the standard deviation for 300 sets of 3, 10, 30, 50, 100, 200, and 500 target and non-target samples. The middle right and lower right plots show similar results for the densities shown in the upper right plot, for which the target and non-target densities have respective mean and standard deviation parameters of 0.393, 0.134, and 0.381, 0.118.

30 non-target samples from each density. Fit beta densities to the target samples and non-target samples. Form a ROC curve from these target and non-target density estimates. Repeat this process many times for many different sets of 30 target samples and 30 non-target samples. The mean of the ROC curves generated approximates the ROC curve of the underlying densities. Similarly, the mean of the AUC values generated from such a process approximates the AUC value of the underlying densities.

Figure 4.12 illustrates results of a process that characterizes the accuracy of AUC values; this process is of interest for characterizing RSD values. First, assume a non-target density. Then, for each target density, find the corresponding AUC value. For the fixed non-target density, the relation of AUC value to the mean and standard deviation of the non-target density is shown in Figure 4.12. The method developed here is still appropriate in the presence of ROC curve or AUC value bias (an analysis of CEG curve and RSD value bias, also included in this section, provides further discussion).

4.2.2 The ROC curve confidence bounds. The explanation here largely focuses on confidence intervals at selected false alarm probabilities, but it extends to confidence intervals over the entire ROC curve, which form confidence contours, and to the confidence band enclosed by the contours. Ideal performance metric confidence intervals may achieve two objectives. First, the stated coverage accuracy of the confidence intervals should be consistent with the actual coverage, where coverage accuracy summarizes actual containment; for example, 90% confidence intervals ideally contain truth with 90% probability. Second, the confidence interval widths should be as small as possible.

The following steps evaluate confidence interval accuracy over a large number of runs.

1. Select a target and a non-target density and find the true score-threshold ROC curve associated with these densities. The true ROC curve is found by evaluating the function

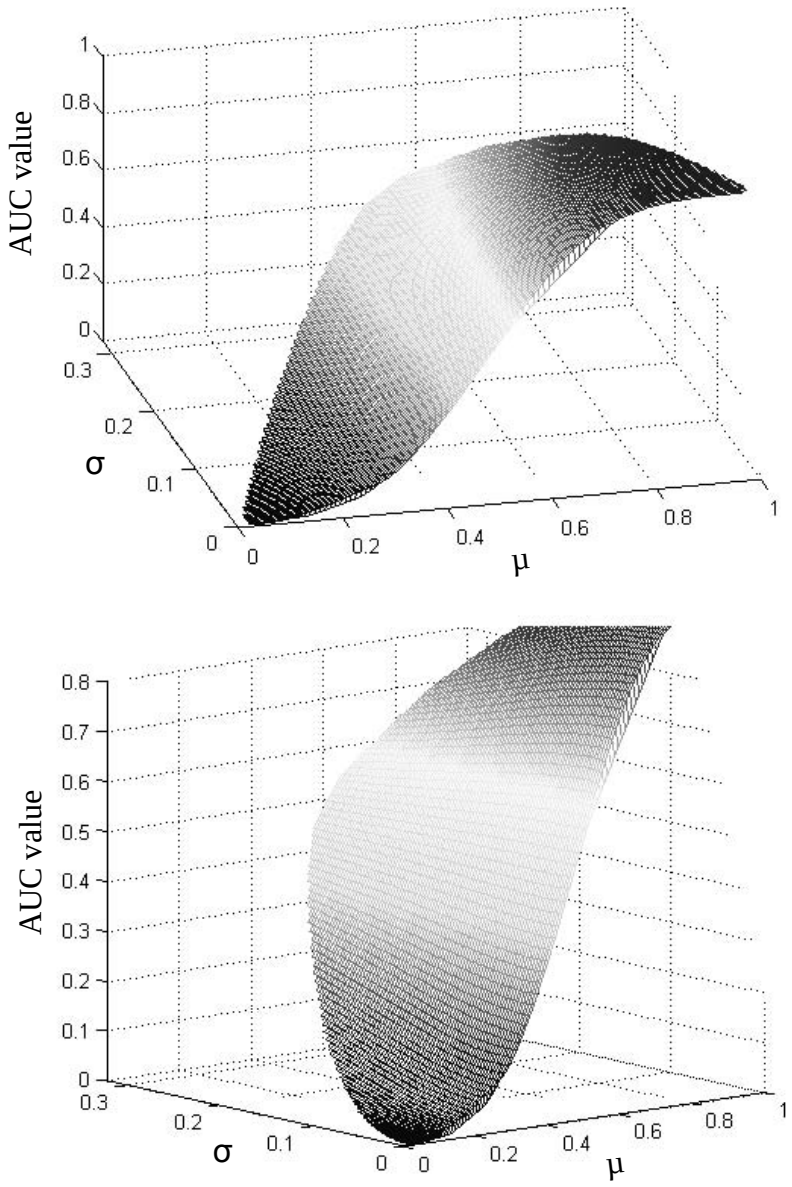


Figure 4.12 Comparison of AUC values for a fixed non-target score density. Here the non-target score density is fixed at $\mu = 0.599$ and $\sigma = 0.021$. The plots show the effect of varying the target density parameters (μ and σ) for the fixed non-target density parameters. The top and bottom plots are the same except for orientation; two plots are provided to facilitate comparison with the RSD value plots of Figure 4.22.

that generates the ROC curve (by varying the score-threshold t) as described in Equations (2.5) to (2.10).

2. Generate many sets of target and non-target score samples from these densities, where each set of samples has the same number of target and non-target samples.

3. Generate for each set confidence intervals for the ROC curve at each of uniformly spaced false alarm probabilities.

4. Record the fraction of instances, called alpha, where the truth (i.e., the true ROC curve) is outside of the confidence intervals; for 90% confidence intervals this fraction is ideally 0.10.

5. Generate a summary alpha value for the entire confidence band by finding the percentage of correct detection probabilities where the confidence intervals do not contain truth for all false alarm probabilities and for all sets.

The Bayesian framework developed here actually produces confidence intervals that reflect coverage probability for particular runs (for the samples, assumed model, and assumed priors); other approaches focus on confidence interval accuracy only over a large number of runs. Note that the steps above are not in themselves concerned with performance for a particular run, and thus these steps perform a frequentist-type verification that evaluates "on average" performance over many runs (or sets of target and non-target samples) (see [Carlin and Louis, 2000, pp. 35-36]). However, it is of interest to test the performance of the Bayesian approach over a large number of runs (as the confidence interval results over one run, although correct, are not possible to verify numerically, except over many runs).

The lower left plot of Figure 4.13 shows that the observed alpha for a particular run can range from 0 to 1. The summary alpha value over all runs for the example shown in Figure 4.13 is 0.09, which approximates the ideal alpha of 0.10.

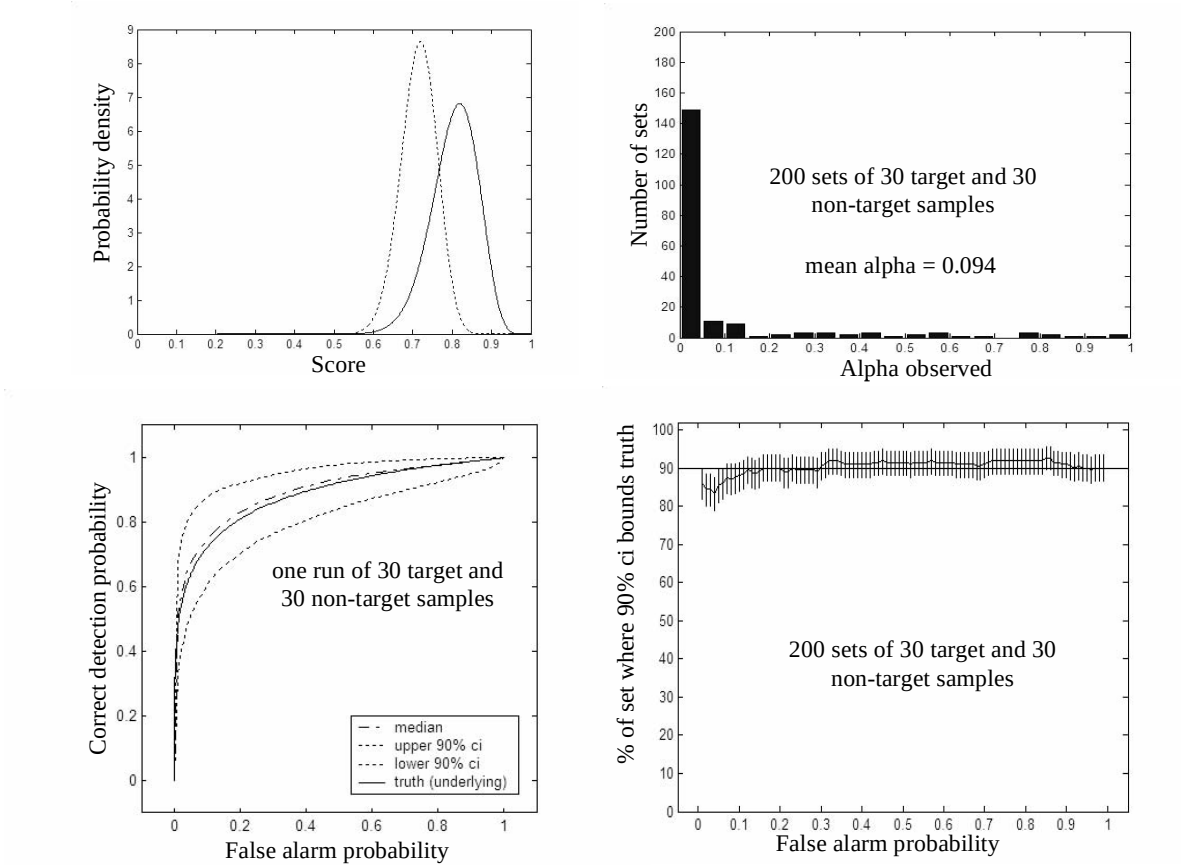


Figure 4.13 Densities, ROC curves, alphas, and coverage for a selected density pair. Here the beta densities of the upper left plot generate 30 target and 30 non-target samples (the densities have $\mu = 0.805$, $\sigma = 0.059$, and $\mu = 0.715$, $\sigma = 0.046$, respectively). The confidence intervals for the ROC curve are shown at the lower left. The upper right plot shows the observed alphas for 200 sets of 30 target and 30 non-target samples, where the mean over many runs should approach 0.10; the observed mean alpha is 0.092. The lower right plot investigates possible bias; results show the process to be unbiased, where vertical lines are 90% confidence bars; these bars narrow as the number of sets increases.

Note that if the summary alpha value described above results in an ideal alpha, the confidence intervals of correct detection probability at a particular false alarm probability are not necessarily ideal. Thus, it is of interest to evaluate the fraction of sets or runs where the separate confidence intervals at particular false alarm probabilities enclose truth. The lower right plot of Figure 4.13 provides an example, where the straight horizontal line indicates ideal 90% coverage and the vertical error bars describe uncertainty due to the finite number of test runs (as the number of test runs increases, the length of each vertical error bar decreases). The coverage of each run is assumed to be from a binomial density; the figure shows 90% vertical error bars based on this assumption. The process described above for developing confidence intervals is optimal for the assumed models, the assumed priors, and the given input samples. Thus any deviation in the coverage accuracy of confidence intervals is due to inapplicable model density forms or inapplicable prior densities of model parameters. Figure 4.14 provides an example for different underlying target and non-target densities.

A similar process is used to develop coverage estimates for AUC value confidence intervals, CEG curve confidence intervals, and RSD value confidence intervals. Figure 4.15 shows the ROC curve density and density contours that corresponds with the confidence intervals of Figure 4.14. Coverage estimates for an AUC value example are shown in Figure 4.16. The upper plot shows the true ROC curve (solid line) and 90% confidence intervals (dashed line) for a single run of an assumed density model. The lower plot shows the AUC value estimate (solid curve) and AUC value 90% confidence intervals (dotted curves) for many separate runs. The calculated alpha value is 0.0993, which approximates the ideal AUC value for 90% confidence intervals.

Attempts to describe coverage accuracy often result in an apparent paradox. For example, assume that 30 target samples and 30 non-target samples are available. Then form ROC curve confidence intervals as detailed in Section 3.4. While this single set of samples forms confidence intervals, coverage accuracy estimation requires many sets of

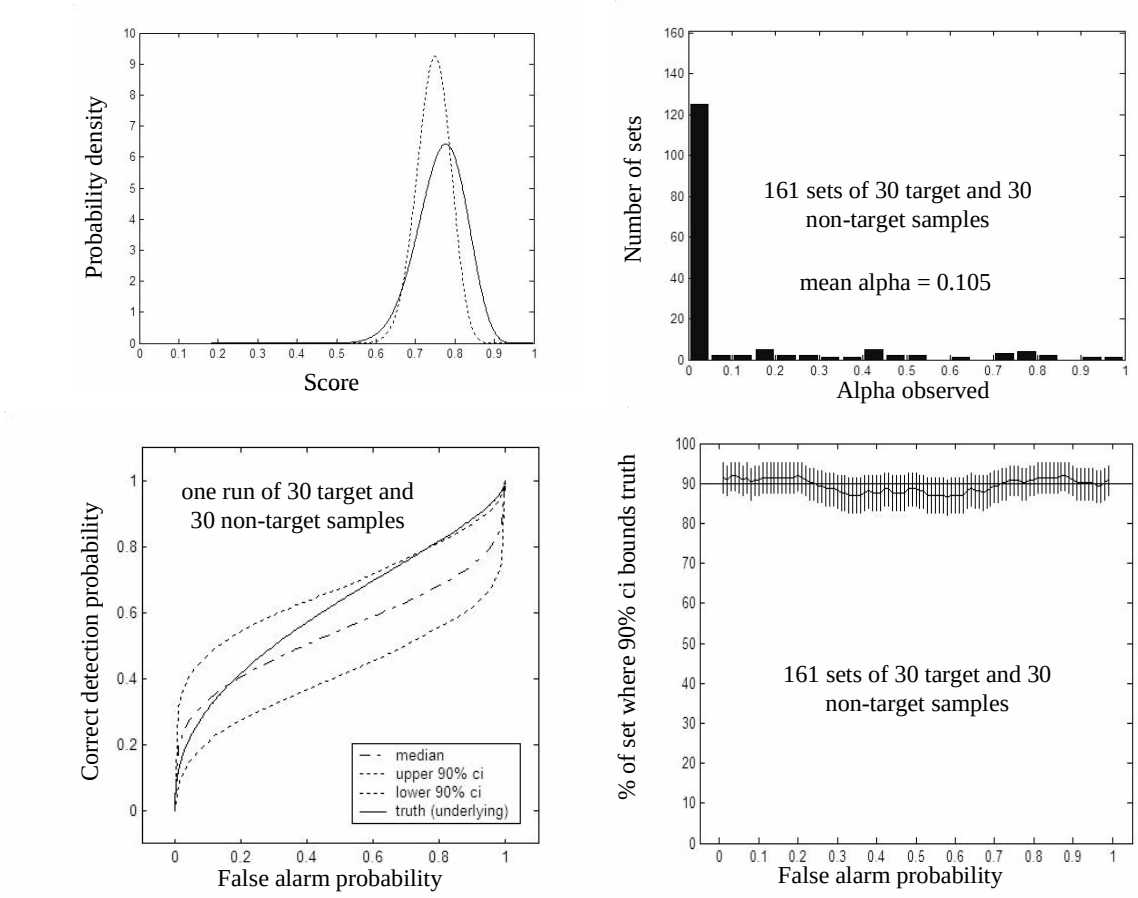


Figure 4.14 Densities, ROC curves, alphas, and coverage for a different target and non-target density pair (these beta densities have $\mu = 0.65$, $\sigma = 0.062$, and $\mu = 0.745$, $\sigma = 0.043$, respectively). This figure repeats the analysis of Figure 4.13 for a different target and non-target density pair.

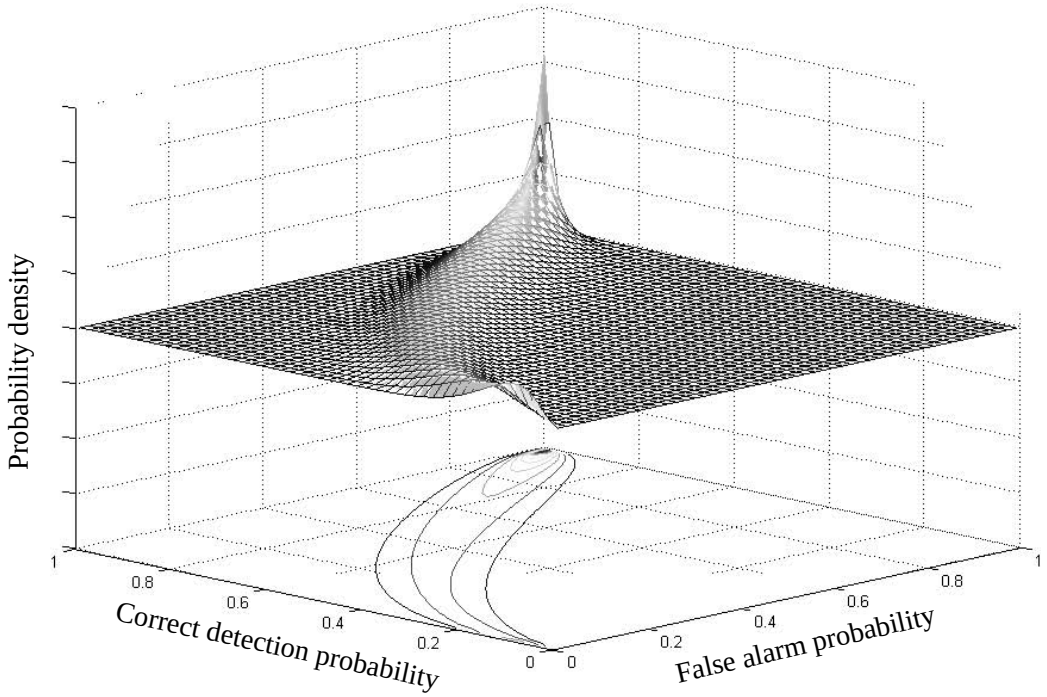


Figure 4.15 A ROC curve density and density contours. The ROC curve density and density contours that correspond with the confidence intervals of Figure 4.14 are shown.

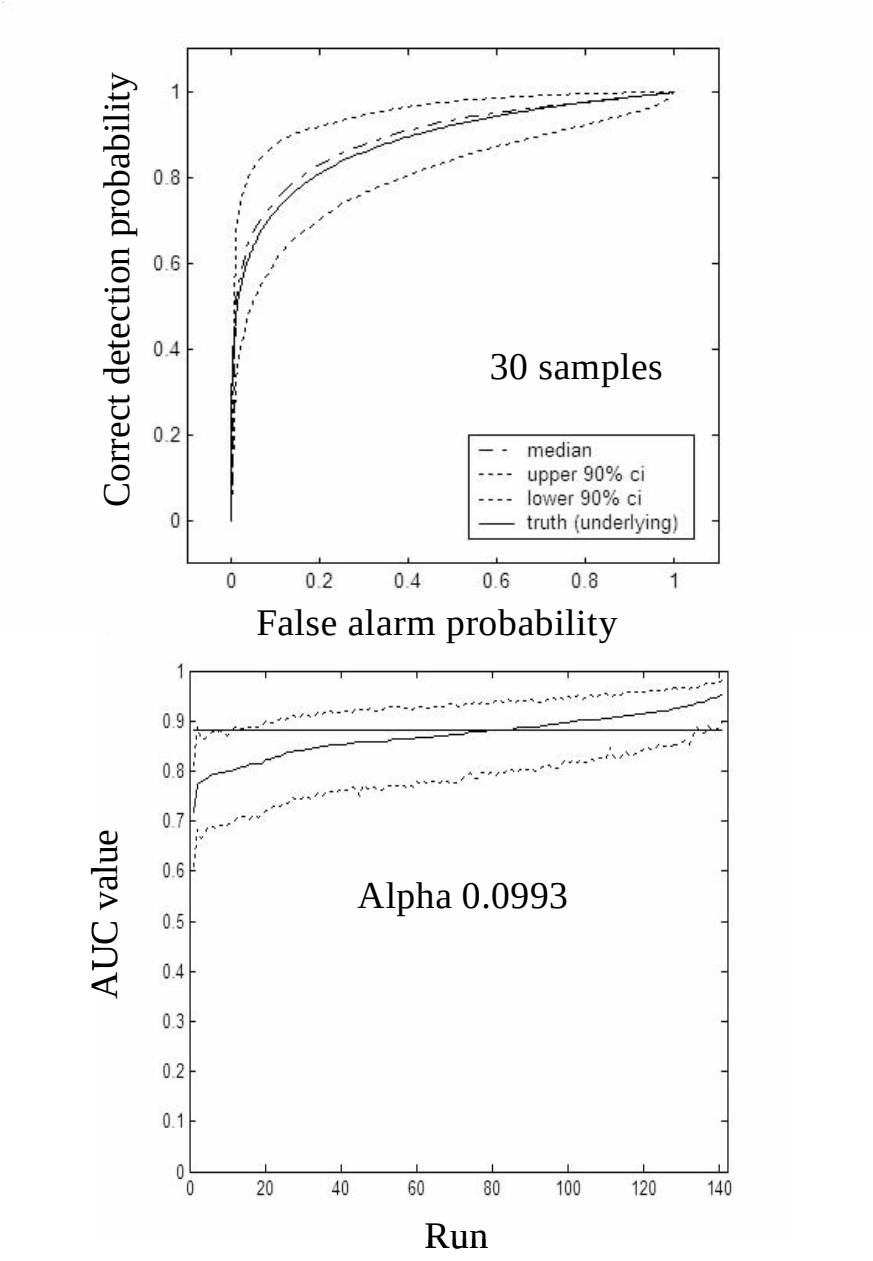


Figure 4.16 Estimates of ROC curves and AUC value confidence intervals. The upper plot shows the true ROC curve (solid line), the median ROC curve (dash-dotted line), and 90% confidence interval contours (dashed lines) for a single run of the density model of the top left plot of Figure 4.13. The lower plot shows the AUC value estimates (solid curve) and AUC value 90% confidence intervals (dotted curves) for many separate runs sorted by lowest to highest estimated AUC value. The straight horizontal line indicates the true AUC value for an infinite number of samples. The calculated alpha value is 0.0993, which approximates the ideal AUC value of 0.10.

target and non-target samples. However, if these sets of samples are available, they may be concatenated so that the number of target samples and non-target samples is much greater than 30. Thus, if enough information is known to test confidence interval accuracy, then enough information is known to make the confidence intervals unnecessary. This discussion identifies a need for representative test data. For such test data either the underlying target and non-target score sample densities are known, or a very large number of target and non-target score samples are known (the latter is the case for the experimental results of the following section).

4.2.3 ROC curve experimental data results. Chapter 3 develops a Bayesian framework that generates performance metric densities. From this framework, various descriptive statistics are derived. The framework and descriptive statistics have in large part been demonstrated with a beta density model; however, they apply to other density models, such as beta mixture models or Gaussian models. An example of this extension is described here using experimental data from an actual SUT rather than data generated from assumed underlying target and non-target densities. The Air Force Research Laboratory (AFRL) made this data available by applying a mean-square/generalized likelihood ratio test (MS /GLRT) algorithm to Moving and Stationary Target Acquisition and Recognition (MSTAR) public data (see [Bryant, 2002]).

Figure 4.17 shows the experimental target and non-target data, after normalization to zero to one. The following procedure selects a full set of target and non-target samples, starting with 588 target scores. The AFRL data has nine sets. Sets 1, 2, and 3 pertain to SAR images that all contain a BMP2 vehicle. Set 1 is the collection of 196 images of a selected BMP2 vehicle. Set 2 is the collection of 196 images of a second BMP2 vehicle. Set 3 is the collection of 196 images of a third BMP2 vehicle. For each of the 586 images in these three sets, an MS/GLRT algorithm has been applied by Bryant to obtain three values [Bryant, 2002]. The first value describes the match of the image to a BMP2, the second value describes the match of the same image to a BTR70 (armored personnel

carrier), and the third value describes the match of the same image to a T-72 (tank). Here, only the first value is of interest (because the BMP2 is assumed to to be a target), and thus 196 x 3 (588) target scores are obtained. Similarly, 784 non-target scores are obtained as follows. Sets 4, 5, 6, and 7 each consist of 196 images that contain a selected BTR70, T-72 (tank 1), T-72 (tank 2), and T-72 (tank 3), respectively. As in sets 1, 2, and 3, the MS/GLRT algorithm has been applied to each set to obtain three values (the match of the image to BMP-2, BTR70, and the T-72). Since the target is the BMP-2, only the first value among the three is retained. Thus there are now 196 x 4 (784) target scores. In addition to the sets of 3-dimensional data values, AFRL provided code that assists in the above process. Note that there are many options for obtaining example target and non-target samples in addition to the method described above. An alternative option takes the three (BMP-2, BTR-70, and T-72) values for each image and retains the highest among the three real number values. In such an alternative, an SUT achieves success as long as it correctly identifies that an image contained a weapon system; the SUT would not necessarily need to identify the specific system.

Sets 8 and 9 are not weapon systems (for example, set 9 contains only bulldozers). Initial normalization ensures that all values within the nine sets of data range from zero to one. Since many of these values are not used when BMP2 is the assigned target, the 588 target scores and 784 non-target scores have a narrower range than zero to one. The lowest value among the 588 target scores and 784 target scores is approximately 0.4 and the highest value is 1. Note that if a score of exactly zero or exactly one is tested in a beta density based model, the posterior density equals zero. Therefore, an additional linear transformation is applied to the data such that all values within the nine sets of data have an upper limit of 0.95 and a lower limit of 0.05.

Here two comparison processes estimate the ROC curve densities and generate ROC curve confidence intervals. The first process applies a single beta density model. The second process applies a two-beta mixture density model. Note that in the two-beta

density model, the number of target and non-target grid points required is large; an exhaustive iterative search over uniform means and uniform standard deviations is not used. Instead, grid points are selected in a uniform, random manner over all allowable means, standard deviations, and ratios, such that an example two-beta density is fully defined by two means, two standard deviations, and one ratio. The ratio shows the relative weighting of the two beta densities that comprise the two-beta density model.

Figures 4.18 and 4.19 show the results. Note that since the underlying densities are not known, the experimental data coverage accuracies and alphas are not expected to be as ideal as in the examples of previous sections. Figure 4.18 assumes a single beta model for the data. Many sets of 30 target samples and 30 non-target samples are drawn from the 588 target scores and 784 non-target scores, and the assumed truth is the ROC curve formed by all 1372 scores.

The figure shows confidence intervals developed for one run of 30 target and 30 non-target samples (drawn from the 588 target scores and 784 non-target scores) and coverage accuracy based on 105 such sets. Note that the ideal mean alpha is 0.1, and the observed alpha is 0.2359. Figure 4.19 applies a two-beta mixture model to the same process. The two beta mixture model has 5 parameters (two means, two standard deviations, and a ratio of the two beta densities). The mean alpha for this bimodal two-beta density mixture model is $0.1038 \simeq 0.10$, which improves the single beta model results.

The lower left plots of both Figure 4.18 and 4.19 show confidence intervals developed by the single beta models and the two-beta mixture models for the same set of 30 target and 30 non-target samples. The upper left plots of the figures use the same set of target and non-target samples, and the plots show the target and non-target densities that correspond to the ROC curve with the highest posterior density or weight (see Figure 3.8). Even though the target and non-target densities of the highest posterior density for the single beta density model do not appear to be of the same form as Figure 4.17, the ROC curve

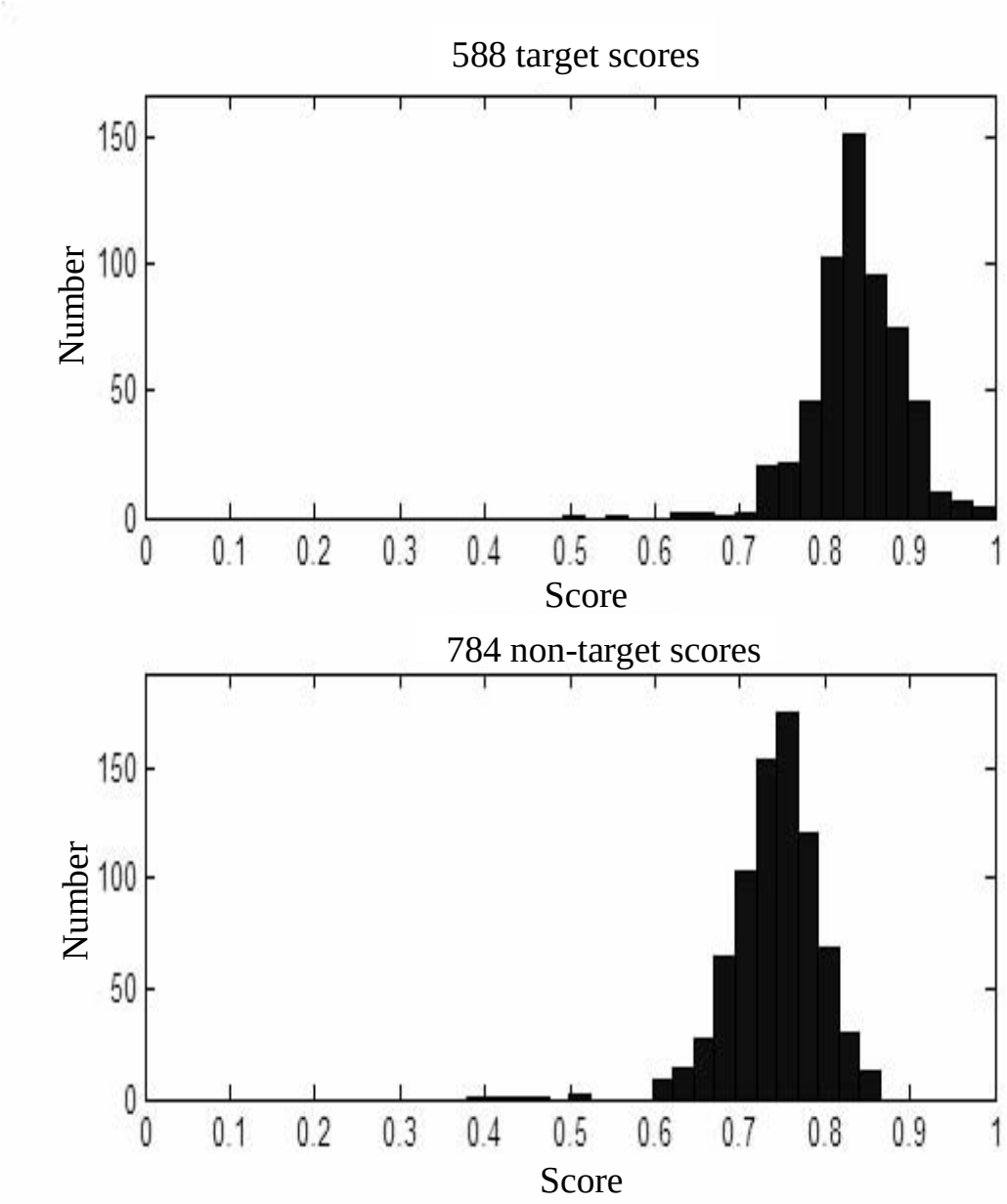


Figure 4.17 Experimental target and non-target score histograms. Based on a subset of data from AFRL/SN [Bryant, 2002], the confidence interval development process (see Figures 4.5 and 4.13) is applied to the experimental data shown above. A single beta density model is applied to this data in Figure 4.18, and a two-beta mixture density model is applied in Figure 4.19. Note that a beta density model requires scaling of the data (since the data here must range from 0 to 1).

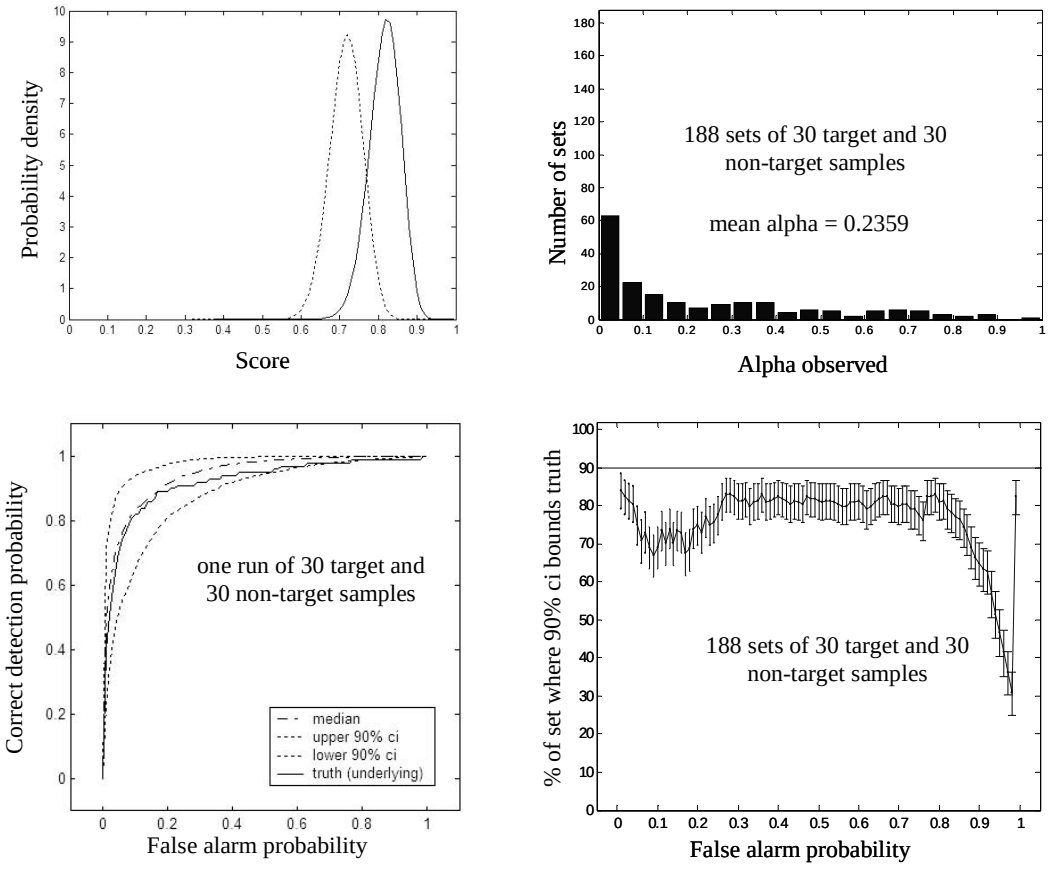


Figure 4.18 Densities, ROC curves, alphas, and coverage for 30 target and 30 non-target samples generated from the experimental data shown in Figure 4.17 and a single beta model. The data of Figure 4.17 is scaled for a maximum range of 0.05 to 0.95 rather than 0 to 1 (see text).

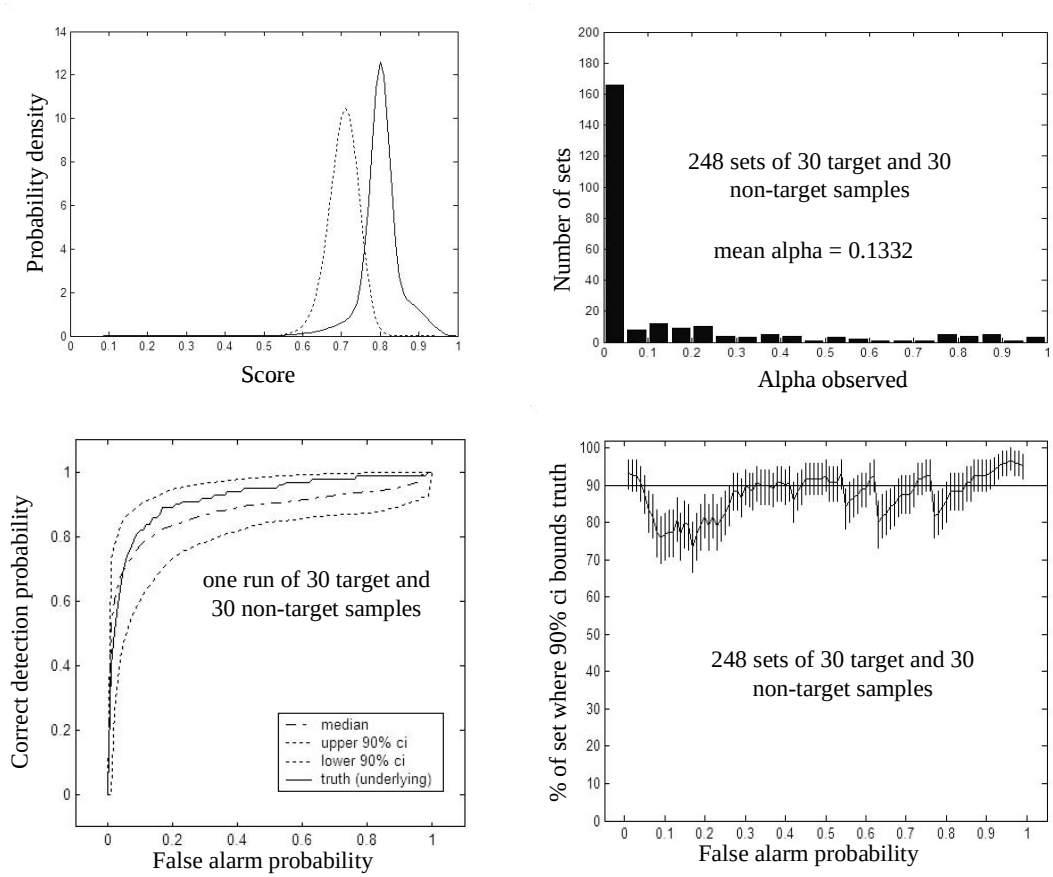


Figure 4.19 Densities, ROC curves, alphas, and coverage for 30 target and 30 non-target samples generated from the experimental data shown in Figure 4.17 and a two beta mixture model.

confidence intervals appear reasonable. This difference emphasizes the benefit of a Bayesian approach. Further, the two-beta mixture model appears to have better coverage accuracy over all false alarm probabilities, showing that more complex models may be of benefit when the density form is not known; such as in this experimental data (for example, note the small but significant number of target samples between 0.5 and 0.7, and the small but significant number of non-target samples between 0.4 and 0.6).

Note that the comparison "truth" is actually an estimate of truth as it consists of only 588 target scores and 784 non-target scores. These numbers seem large enough to approximate truth, but there is uncertainty (see Figures 3.2 and 3.3, and related discussion). This result also emphasizes the importance of incorporating knowledge of the actual underlying model, if known. MacKay [MacKay, 2003] discusses the related concept of importance sampling, which provides the option of using a simpler model even when it is known that a more complex model is truth.

Additional implementation choices exist. An option is to change the scaling of the data. If the data were scaled from 0.1 to 0.9 rather than 0.05 to 0.95, the scaling may impact coverage accuracy. An example for the single beta density case for 0.1 to 0.9 scaling is shown in Figure 4.20. For this example, the change in scaling has minimal impact on the results.

The results presented here show the ability of the framework to evaluate experimental data. This results presented here do not imply that the two-beta density mixture model will always have results that improve a single beta model. The single beta density framework (and two-beta density mixture model extension) have been introduced in this research as examples to test the framework developed in Chapter 3. Detailed approaches regarding the appropriate incorporation of more complex models are presented in future work.

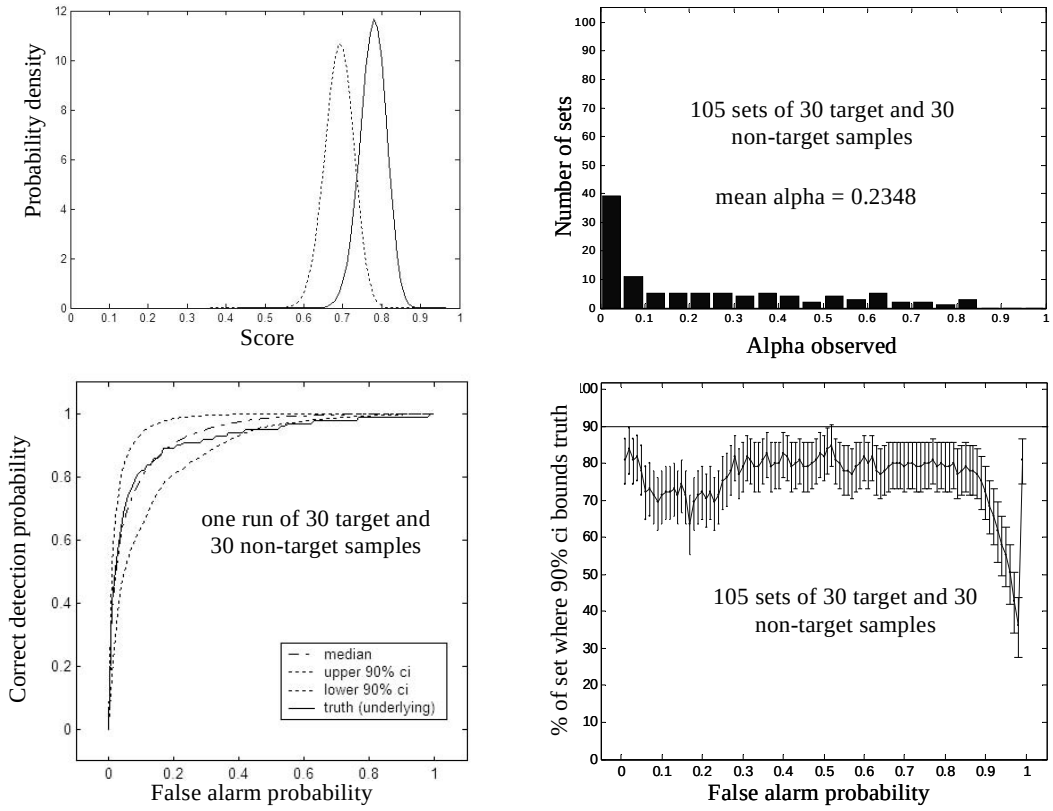


Figure 4.20 Same as Figure 4.18, except that the experimental sample values are scaled for a maximum range of 0.1 to 0.9.

4.2.4 Analysis of CEG curve and RSD value bias. The results that follow in Section 4.2.5 quantify CEG curve confidence band accuracy by repeated runs over many sets of samples. Before examining this accuracy, consider that RSD values formed by fitting beta densities to beta density generated data can have a higher bias than AUC values, particularly for low numbers of samples. For example (see Figure 4.21), select a target and non-target beta density pair. Generate 30 target and 30 non-target samples from each density. Fit beta densities to the 30 target and 30 non-target samples by matching sample and density mean and variance. Form a CEG curve and RSD value from these two beta density estimates. Repeat this process many times for many different sets of 30 target samples and 30 non-target samples. The mean RSD value generated from this process may be consistent with the RSD value of the underlying densities. Note that the CEG curve estimates exhibit a slight bias, but the standard deviation is wide.

In Figure 4.22 a non-target density is assumed, then the RSD value is found for many target beta densities. If truth is at the minimum of the "bowl" shown, then the verification process that was used for AUC values is not appropriate for RSD values (compare Figure 4.22 with Figure 4.12). However, RSD values developed here are appropriate: given an assumed model of beta densities for target and non-target and given target and non-target samples, 90% correct confidence intervals for RSD values can be generated. These confidence intervals are correct, although they may not enclose the truth for 90% of runs.

The verification issue noted here may be illustrated as follows. Suppose 1000 students take a test of 100 questions. It is known (as a prior) that 999 of the students answer 80 questions correctly and one student answers 95 questions correctly. An evaluator is aware of this information and obtains 10 test questions from a randomly selected student. Unknown to the evaluator, the selected student is the student who answers 95 questions correctly. The evaluator is to provide 90% confidence intervals for the number of questions that the student answers correctly. Based on the priors, the evaluator specifies the upper and lower 90% confidence intervals at 80 questions correct. This process is

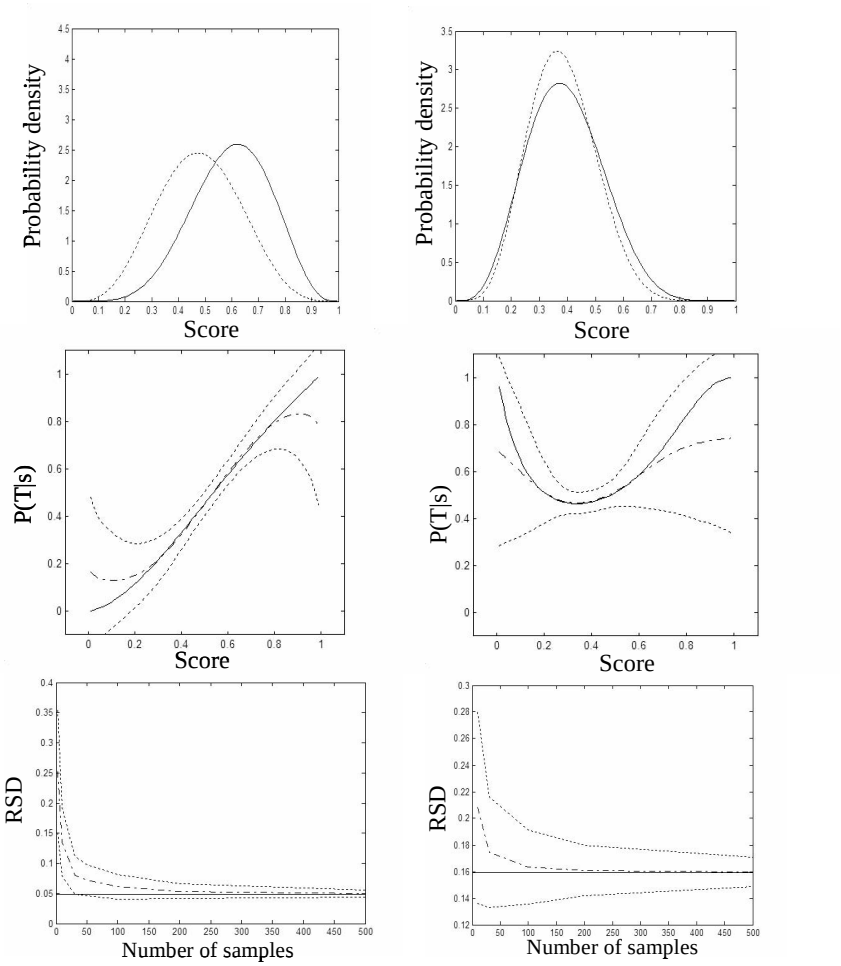


Figure 4.21 Estimates of CEG curves and RSD values. The top two plots show underlying target densities (solid curves) and underlying non-target densities (dashed curves). The middle left plot shows the CEG curves for the underlying beta densities (solid curve) with CEG curve statistics for 300 sets of 30 target and 30 non-target samples drawn from each density shown in the top left plot, where the mean of the 300 curves (dash/dotted line) and this mean plus and minus the standard deviations are plotted (dotted lines). The lower left plot similarly shows the true RSD value, mean RSD value, and mean RSD value plus and minus the standard deviation for 300 sets of 3, 10, 30, 50, 100, 200, and 500 target and non-target samples. The middle right and lower right plots show similar plots for the densities in the upper right plot.

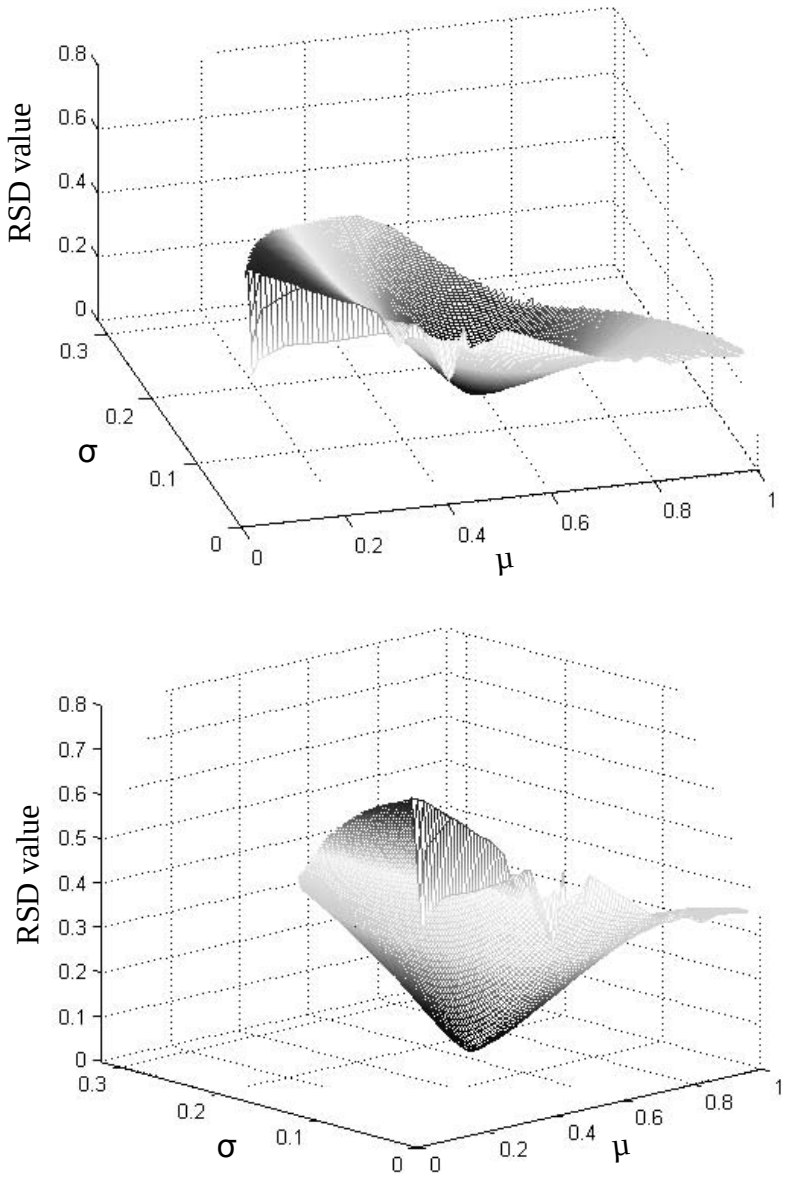


Figure 4.22 The RSD values for a fixed non-target density. Here the non-target density is constant and the target density varies over the full range of possible beta parameters. Note the bowl appearance, where RSD approaches a minimum at mean of 0.6 and standard deviation of 0.1. If the true density has the minimum RSD value, then the RSD confidence intervals developed for small set of samples do not enclose truth because uniform priors over mean and standard deviation are assumed. The confidence intervals are reasonable even though they are not necessarily appropriate in the standard coverage accuracy test used for the CEG curve, ROC curve, and AUC value confidence intervals. The two plots are the same except for orientation.

repeated many times. No matter how many sets of 10 questions are provided, when each set is considered individually, the confidence intervals will never enclose the truth of "95 questions correct".

4.2.5 The CEG curve confidence bounds. Figure 4.23 is similar to Figure 4.13, except that the performance metric examined is the CEG curve rather than the ROC curve. Using the accuracy description of alpha, as with the ROC curve, CEG curve confidence interval development is shown to be accurate for the assumed model and priors. Note that this figure is representative of CEG curve results; similar plots with sample sizes of 10, 30, 100, and 200 have been tested with similar results (and with an additional underlying density for which the CEG curve is near the 45 degree line). The results are significant because whereas the ROC curve confidence interval process described here is an improvement over existing techniques, the CEG curve confidence interval specification process is without precedent. The results also demonstrate the general extensibility of the entire Bayesian framework to performance metrics other than the ROC curve. Figure 4.24 shows an additional example using different underlying target and non-target densities.

The verification processes that are applied here in Chapter 4 will be used to assist in comparisons with the literature in Chapter 5.

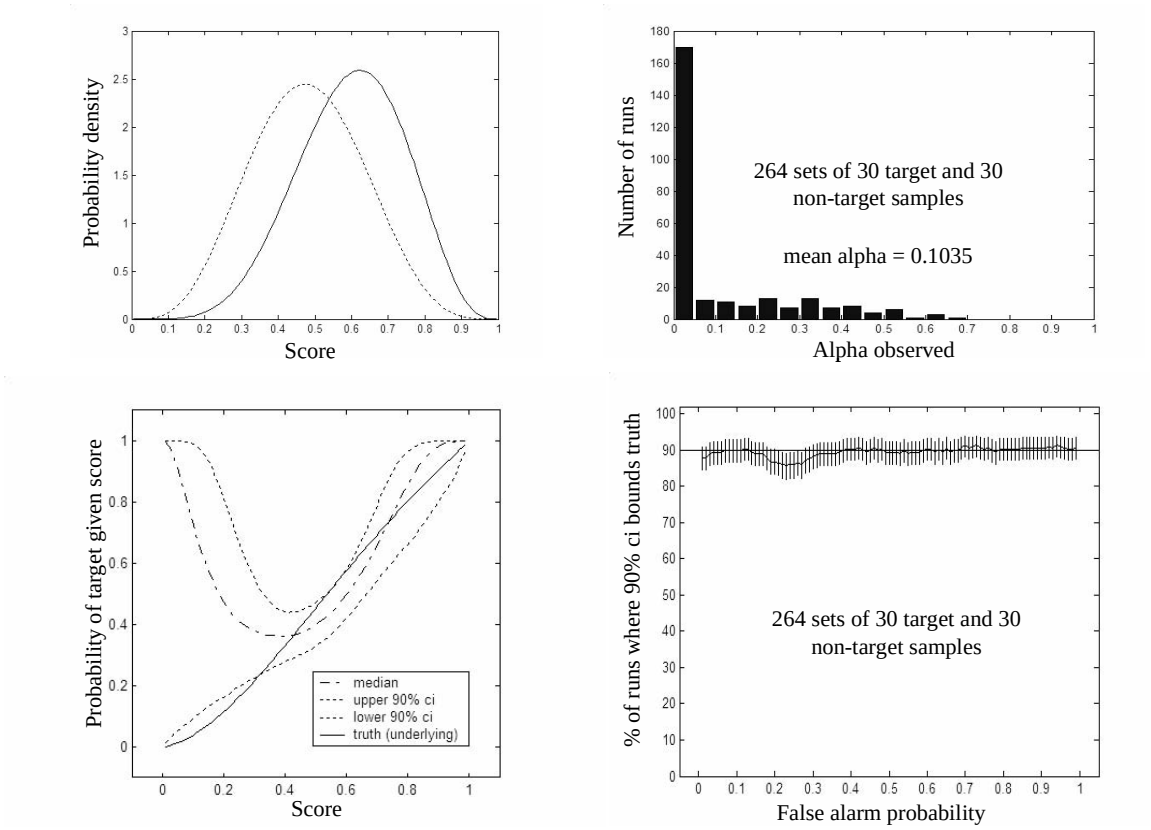


Figure 4.23 The alpha metric for a CEG curve. Here the underlying densities shown at the upper left generate 30 target and 30 non-target samples (the beta densities have $\mu = 0.599$, $\sigma = 0.021$, and $\mu = 0.479$, $\sigma = 0.023$, respectively). Confidence intervals for the corresponding CEG curve are shown in the lower left with the median CEG curve and the true CEG curve. The upper right plot shows the observed alphas for 264 sets of 30 target and 30 non-target samples, where the mean over many runs should approach 0.10; the observed mean alpha is 0.1035. The lower right plot investigates possible bias; results show the process to be unbiased, where vertical lines are 90% confidence bars; these bars narrow as the number of sets increases.

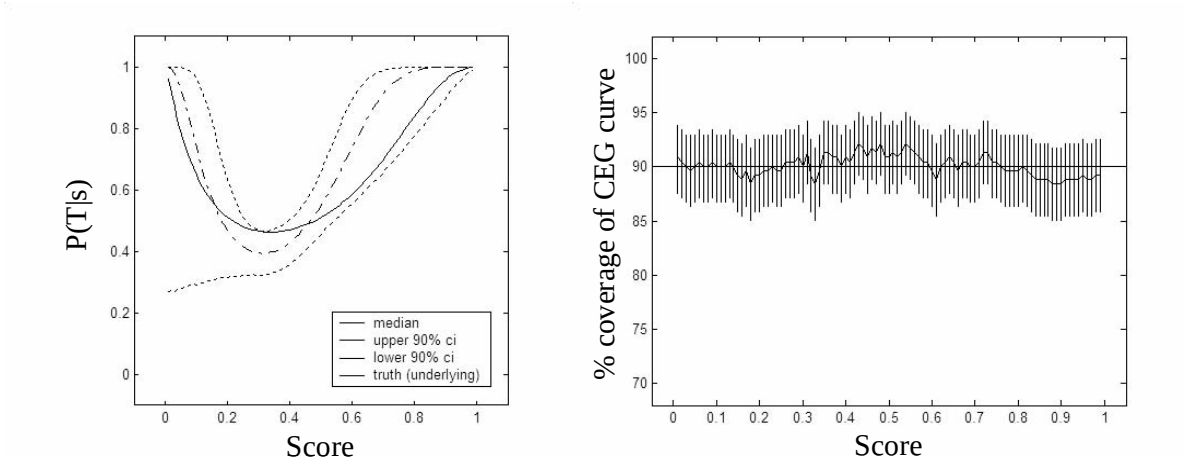


Figure 4.24 The CEG curve confidence intervals for a single run and coverage accuracy over many runs. The left plot shows 90% confidence intervals developed for 30 target samples and 30 non-target samples (the underlying densities are the same as in the right plots of Figures 4.11 and 4.21). The right plot shows the percent coverage of confidence intervals produced for 247 runs, where each run repeats the process used to generate the left plot.

5. Quantitative Comparisons

In this chapter quantitative comparisons are made with methods described in the literature review of Chapter 2. First, the Metz method, which was discussed extensively in the first part of the literature review section of Chapter 2, is now reviewed and compared qualitatively and quantitatively with the method developed here. Then other methods are also reviewed and compared. Here, coverage accuracy and alpha (as described in Chapter 4) are used to quantify the accuracy of the confidence intervals of the method developed here with other available methods in the literature. These metrics provide tools for comparing the accuracy of the developed confidence intervals among various ROC uncertainty estimation methods.

5.1 Comparison with Metz confidence interval method

Figure 5.1 compares the Metz method [Metz *et al.,* 1998] with the method developed here. This evaluation uses the software package ROCKIT to execute the Metz method. Beta densities generate 30 target and 30 non-target samples. Many runs repeat this sample generation process, where each run selects a new set of 30 target and 30 non-target samples. Application of the confidence interval calculation method developed here (see Section 4.1.4) generates unique ROC curve confidence intervals for each run. Confidence band coverage area evaluation and alpha (coverage accuracy) evaluation reveal clear advantages of the method developed here over the Metz method. For many runs of 30 target and 30 non-target samples, the coverage accuracy may be evaluated and averaged over all false alarm probabilities. For 120 such runs, the method developed here is 51% closer to the ideal alpha of 0.05 (for 95% confidence intervals) over the range of the ROC curve. Recall that larger confidence band coverage area without improved coverage accuracy implies less useful results. Again analyzing the 120 repeated runs of 30 target and 30 non-target samples, the Metz method has 16% larger confidence band

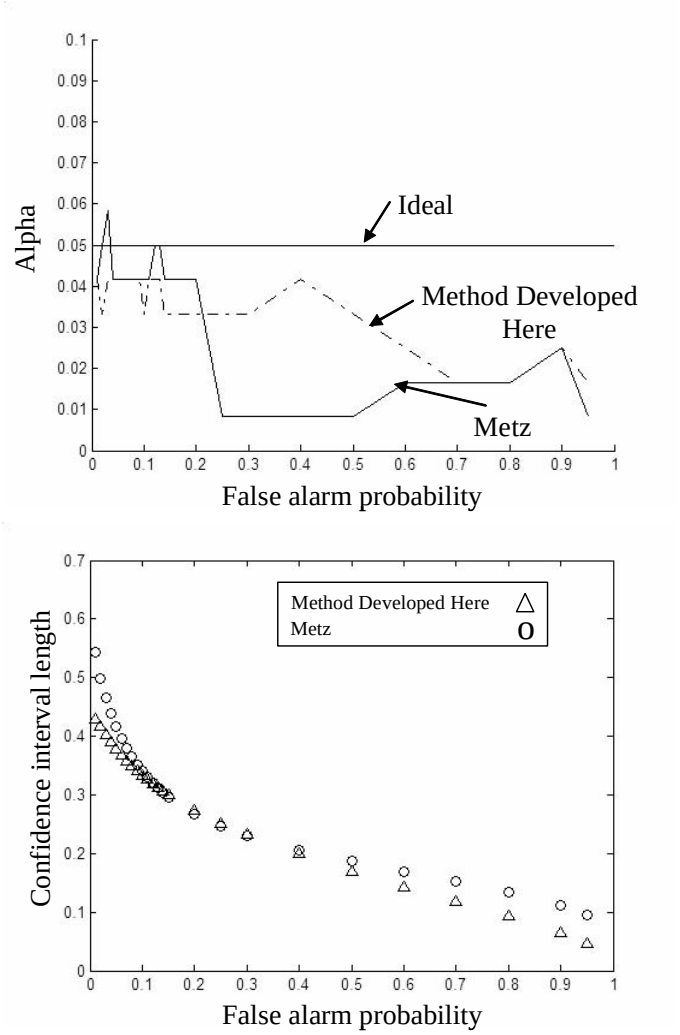


Figure 5.1 Alpha and confidence interval lengths for the Metz [Metz *et al.*, 1998] method and the method developed here. Both methods develop 95% confidence intervals. Beta densities with target mean of 0.805, target standard deviation of 0.059, non-target mean of 0.715, and non-target standard deviation of 0.805 generate 30 target samples and 30 non-target samples many times. Note that the Metz method appears to be slightly closer to the ideal alpha than the method developed here between false alarm probability values of 0 and 0.02 and 0.11 and 0.13, which is not necessarily advantageous because the method developed here has greater coverage (approximately 97%) combined with significantly shorter interval lengths (21% shorter at a false alarm probability of 0.01, for example) at these values. A similar argument applies for false alarm probability values between 0.25 and 0.4, as the confidence interval lengths of the two methods are nearly identical, and the Metz method has wider coverage. For the smallest possible confidence interval widths that maintain at least (1-alpha) coverage, the method developed here outperforms the Metz method for every false alarm probability.

area than the approach developed here, and Metz has larger coverage for the full range of critical false alarm probability values between 0 and 0.2. For the smallest possible confidence interval widths with at least (1-alpha) coverage, the method developed here outperforms Metz at every false alarm probability. Note that in contrast to the Metz method, the method developed here requires no assumptions about the shape of the ROC curve, which is important because for target detection system evaluation it is not appropriate to presuppose the shape. Comparing the top and bottom plots of Figure 5.1, note that there is a false alarm probability (near 0.2), where the Metz method has a higher alpha than the method developed here, but also has a larger confidence interval length than the method developed here. These results are reasonable because confidence interval length does not indicate whether or not the length is over the appropriate range of correct detection probabilities.

The Metz method does not allow for ready incorporation of prior assumptions to refine the ROC curve uncertainty estimates. The choice of a generally convex ROC curve (if only unintentionally) becomes a choice of a prior. Some adjustment or weighting of the covariance terms of the binormal approach could change the standard error, but Metz does not discuss such adjustment. The method developed here permits the ready incorporation of target and non-target parameter priors, and it may be easily extended to any density form.

Figure 5.2 shows the ROC curve and example associated confidence bands for the example of Figure 5.1. Figure 1.3 has already revealed that the confidence intervals for the Metz approach can result in a significantly larger confidence band area than the confidence band area for the method developed here.

Comparison with the Metz method makes clear significant weaknesses in the ability of the Metz method to adapt to curve forms that are not concave. This comparison shows that the Metz method is inferior in confidence interval coverage accuracy and confidence band area compared with the method developed here. However, even disregarding these

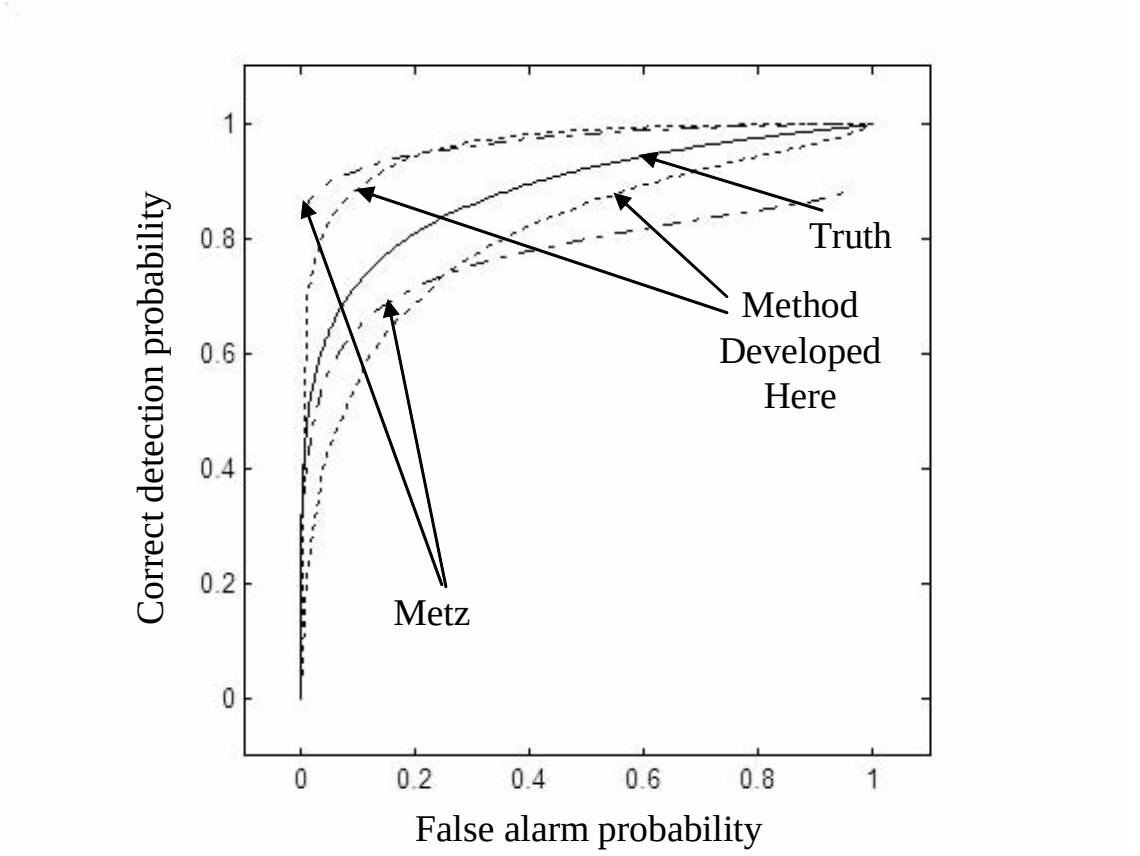


Figure 5.2 Comparison of ROC curve and confidence intervals. Here 30 target and 30 non-target samples are drawn from beta densities for which the solid curve is the true ROC curve for an infinite set of samples (the target mean is 0.715, the target standard deviation is 0.01, the non-target mean is 0.715, and the non-target standard deviation is 0.046). The 90% confidence interval contours for the method developed here and the Metz method are shown. Figure 5.1 reports the coverage accuracy and confidence interval widths for many runs, and the plot shown here gives one example of such a run.

disadvantages, the Metz approach does not apply to the confidence error generation (CEG) curve or other performance metrics where the assumed form of the performance metric curve is not a straight line in normal deviate space.

5.2 Comparison with Zhou confidence interval method

The literature considers various ROC-curve bootstrap approaches, i.e., methods that generate confidence bounds using subsets of the available target and non-target samples. This section examines the most recent approach, Zhou [Zhou and Qin, 2005], who obtains results that improve upon the bootstrap results of Platt [Platt *et al.*, 2000]. A general advantage for bootstrap methods is that they make no assumptions about the form of the densities (such as assuming a beta density). Both Platt and Zhou claim reasonable coverage accuracies for 95% confidence intervals of correct detection probability at false alarm probabilities of 0.1 and 0.2; Zhou claims smaller confidence interval widths.

Zhou develops two new bootstrap-based approaches; the approach that Zhou regards as optimal is used here for comparison. In discussing Platt's work, Zhou points out disadvantages of bootstrap methods, such as the high number of target and non-target samples necessary for accurate results. Zhou claims that a binomial correction factor improves bootstrap-based results, particularly at low numbers of samples. He considers multiple examples with 20 target samples and 20 non-target samples, whereas Platt's research focuses on 100 target samples and 100 non-target samples.

Zhou's paper only considers results at false alarm probabilities of 0.1 and 0.2. Figure 5.3, which corresponds with Zhou's example 2 and 3, uses Zhou's method but develops confidence intervals for other false alarm probabilities. At false alarm probabilities of 0.1 and 0.2, coverage accuracies similar to Zhou's results are obtained (see the top right plot of Figure 5.3). The confidence interval widths are also consistent with Zhou's findings. As Zhou and Platt both focus only on false alarm probabilities of 0.1 and 0.2, a key concern is whether or not confidence intervals are accurate over other false alarm

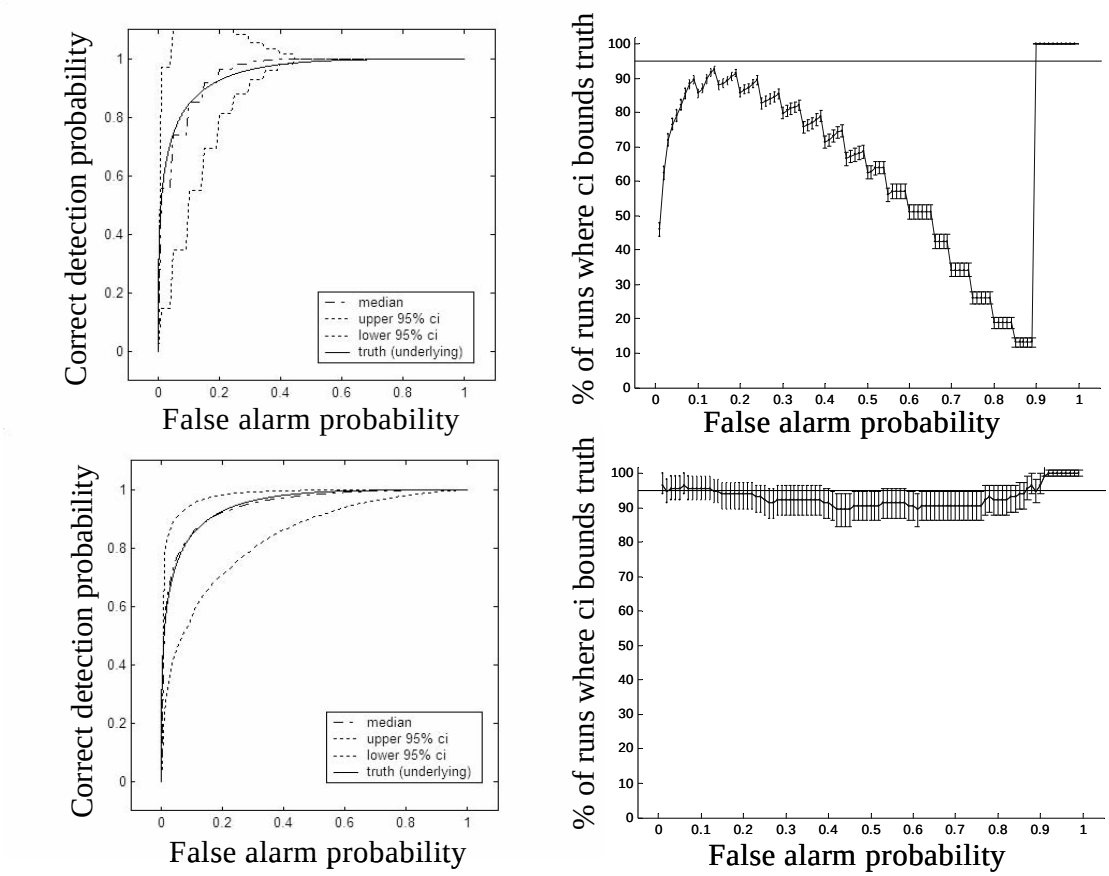


Figure 5.3 Confidence intervals for one run of the Zhou [Zhou and Qin, 2005] method, coverage accuracy for many runs, and comparisons with the method developed here. Zhou examines false alarm probabilities of 0.1 and 0.2, and his work is extended here to the full range of false alarm probabilities. The top left plot shows a representative ROC curve with confidence intervals for the Zhou approach with 20 target and 20 non-target samples. The top right plot shows the percent coverage of the Zhou method for 1700 runs with 90% coverage vertical confidence bars. The lower two plots compare 116 runs for the method developed here. Note that in contrast to Zhou, the method developed here produces smooth confidence bands and ROC curves, and the coverage is consistent over the full range of false alarm probabilities.

probabilities. Examining the top right plot of Figure 5.3, the ROC curve for a density pair that Zhou selects deviates considerably from the ideal 95% coverage between a false alarm probability of 0.3 and 0.88. Recall that bootstrap methods rely only on the observed samples (rather than estimates of densities). If the underlying density that generates the samples is relatively small at a particular score, the corresponding correct detection probabilities at that score are difficult to estimate with a bootstrap method. Figures 5.4, 5.5, and 5.6 show the underlying densities that Zhou uses as examples, results of the Zhou method, and a comparison with the method developed here. In contrast to Zhou and other bootstrap methods, the method developed here has appropriate coverage accuracy over the entire range of the ROC curve.

5.3 Comparison with Hall confidence interval method

Hall [Hall *et al.*, 2004] uses a kernel-based approach to form confidence intervals. They use an updated bandwidth calculation approach that extends previous kernel-based approaches. The method they develop requires use of 10 different smoothing parameters to set different bandwidths. They report coverage accuracy results where samples are generated repeatedly from assumed underlying densities and report results for 100 target samples and 100 non-target samples. These results appear to be generally accurate, except at the extremes of false alarm probability, where the coverage accuracy often declines. This result is of concern, as very low false alarm probabilities are often of particular interest; however, adequate coverage accuracy over the full range of false alarm probabilities is important as indicated, for example, in the SUT A and SUT B example of Chapter 1. (Of course, if it is known a priori that the only false alarm probability of interest is a false alarm probability of 0.5, then the Hall method performs well for the examples reported by Hall.) Figures 5.7 and 5.8 show a comparison of the method developed here and two of the Hall examples. The weaknesses that the Hall method can have at the extremes of false alarm probabilities is apparent in the Figures.

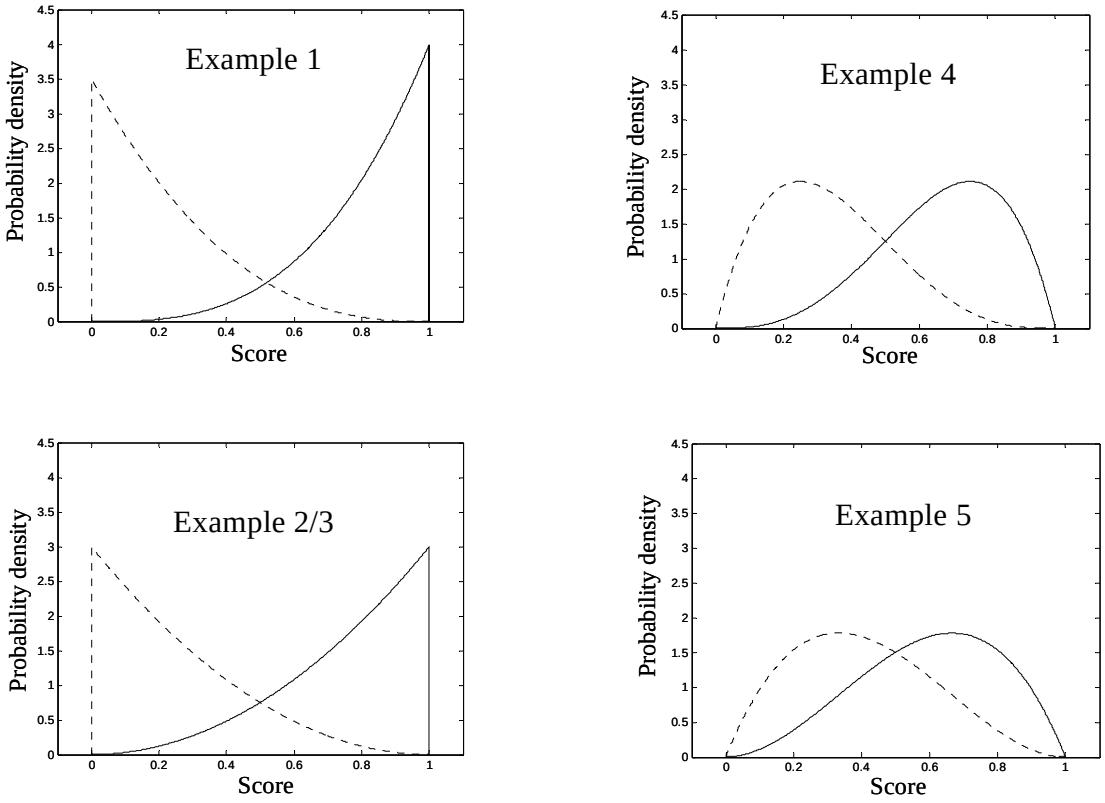


Figure 5.4 Underlying densities for examples used to compare with the Zhou [Zhou and Qin, 2005] method. Zhou selects the above beta densities, and these densities generate target and non-target samples. The solid lines are target and the dotted lines are non-target. Note that examples 2 and 3 are combined because Zhou uses the same underlying densities for two examples (they examine false alarm probabilities of 0.1 and 0.2).

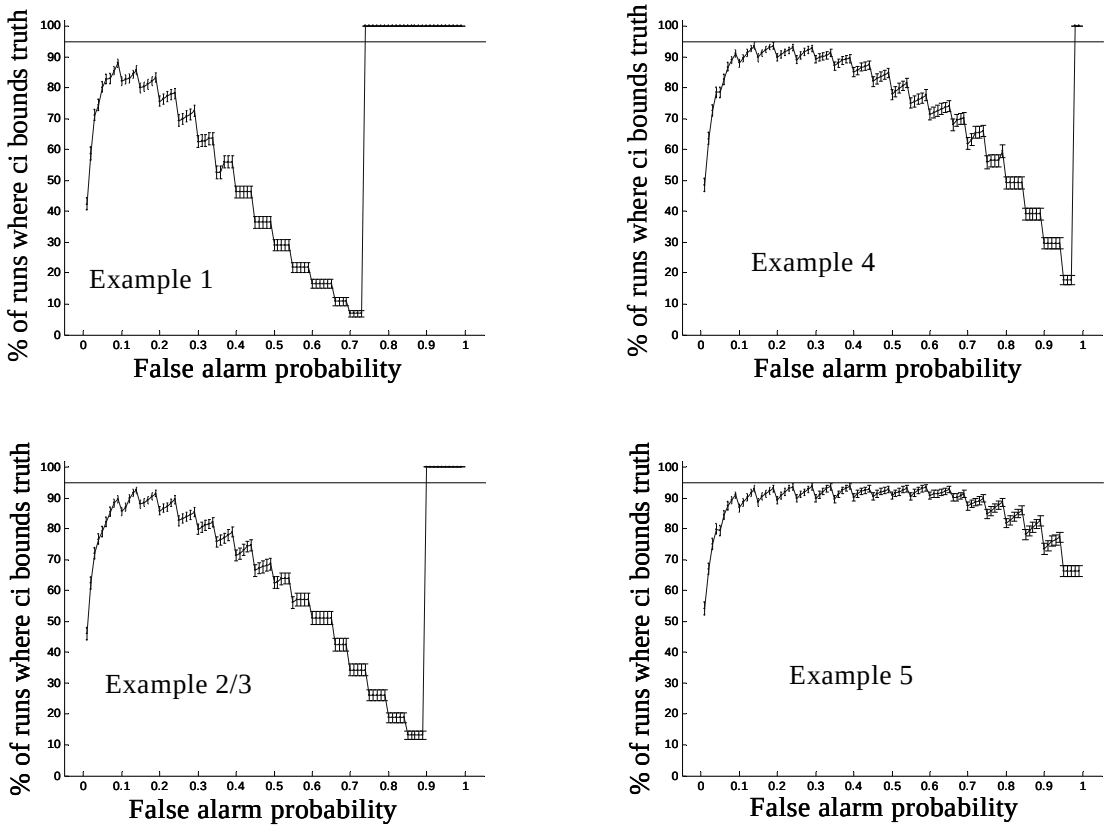


Figure 5.5 Coverage accuracy for Zhou [Zhou and Qin, 2005] confidence bounds. The plots show the percent coverage of confidence bounds for each of the four density pairs of Figure 5.4. Note that Zhou only examines false alarm probabilities of 0.1 and/or 0.2, so examples 2 and 3 have identical underlying densities. These plots are similar to the top right plot of Figure 5.3, except that three additional examples are shown, where 1700 sets of 20 target samples and 20 non-target samples are the inputs.

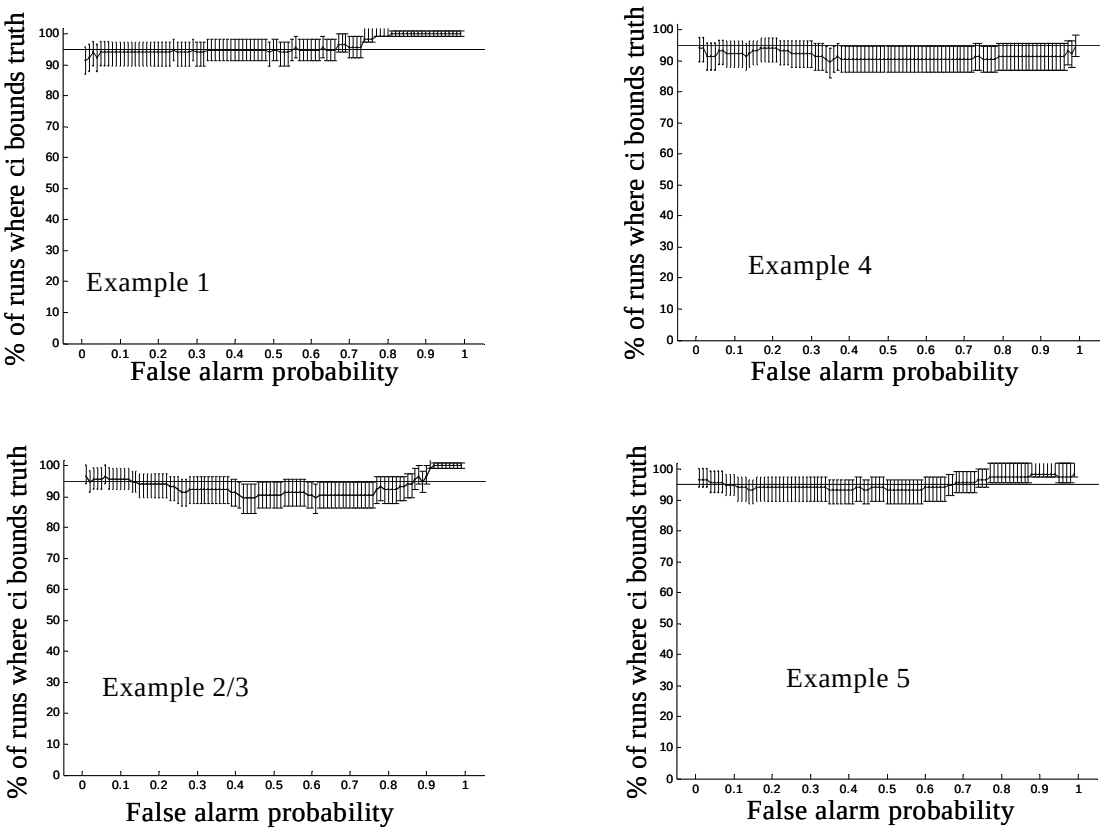


Figure 5.6 Percent coverage of comparison bounds for the method developed here. The plots show the percent coverage of confidence bounds for each of the four density pairs of Figure 5.4 using the method developed here based upon sets of 20 target and 20 non-target samples. Zhou considers only false alarm probabilities of 0.1 and/or 0.2. These plots are similar to Figure 5.3, except that three additional examples are shown.

Figure 5.7 uses normal target and non-target densities, and Figure 5.8 shows beta target and non-target densities. For the method developed here, the normal target and non-target densities first generate samples, then these samples are transformed so that the greatest value among the target and non-target samples is 0.95 and the lowest value among the target and non-target samples is 0.05. In addition to comparing favorably with the Hall approach, the example of Figure 5.7 indicates that the method developed here is flexible to changes in assumed densities.

5.4 Comparison with Hilgers confidence interval method

Figure 5.9 shows confidence intervals based on the Hilgers [Hilgers, 1991] binomial method. The Hilgers method is similar to the current AFRL ROC curve confidence interval estimation approach. The coverages (95% is the objective in the above case) tend to be too conservative, and the resulting confidence intervals are too wide (see discussion in [Schafer, 1994]). The method developed here provides a smoother estimate of the ROC curve (dash/dotted line) than the Hilgers method, and more significantly it produces much narrower confidence intervals, particularly for low numbers of samples.

The Hilgers approach uses a binomial-based ordered statistics approach and finds 95% error bars in correct detection probability and false alarm probability at a selected threshold. The resulting rectangular region then combines two error bars using the following procedure. First, it finds a best-case upper confidence band point for this threshold as the minimum false alarm probability and maximum correct detection probability within the rectangular region. Second, it finds a worst-case lower confidence band point for this threshold as the maximum false alarm probability and minimum correct detection probability within the region. Finally, it repeats for all thresholds and combines results to obtain a lower confidence interval contour and an upper confidence interval contour. This process generates a 95% ROC curve confidence band. Although bands obtained by this process enclose at least 95% of the true ROC curve, the bands are

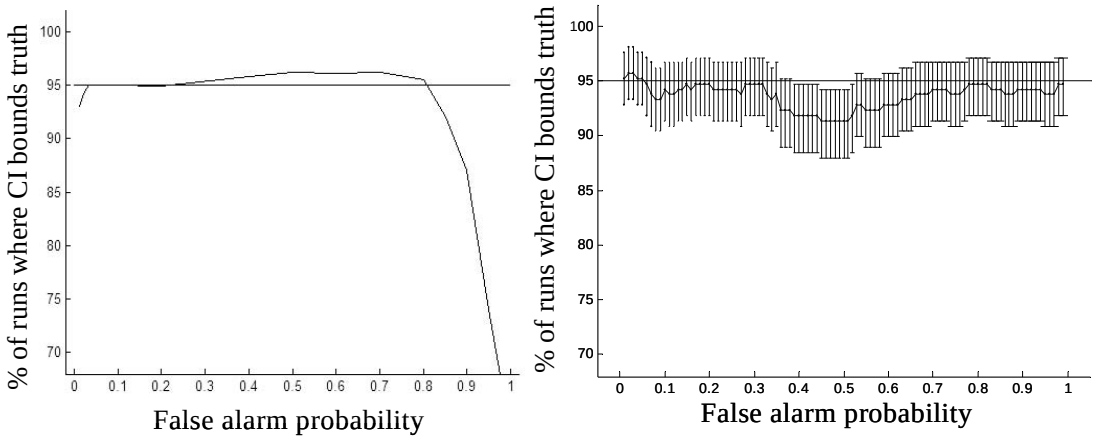


Figure 5.7 The ROC curve confidence interval coverage accuracies for the Hall [Hall *et al.*, 2004] method and the method developed here for normal target and non-target densities. Normal target and non-target densities generate 100 target samples and 100 non-target samples. This process is repeated many times to determine coverage accuracy. The target density has a mean of one, the non-target density has mean of zero, and both densities have unit variance. The plot at left shows Hall's coverage accuracy at selected false alarm probability for 1000 sets of samples. The plot at right shows a similar graph for the method developed here, with 90% vertical confidence bars for 208 sets of samples (90% vertical bars show uncertainty due to the lower number of runs). Hall's coverage accuracy is generally accurate, except as false alarm probability approaches zero or one. This inaccuracy is a weakness in the Hall approach, because often the most significant false alarm probabilities are near zero.

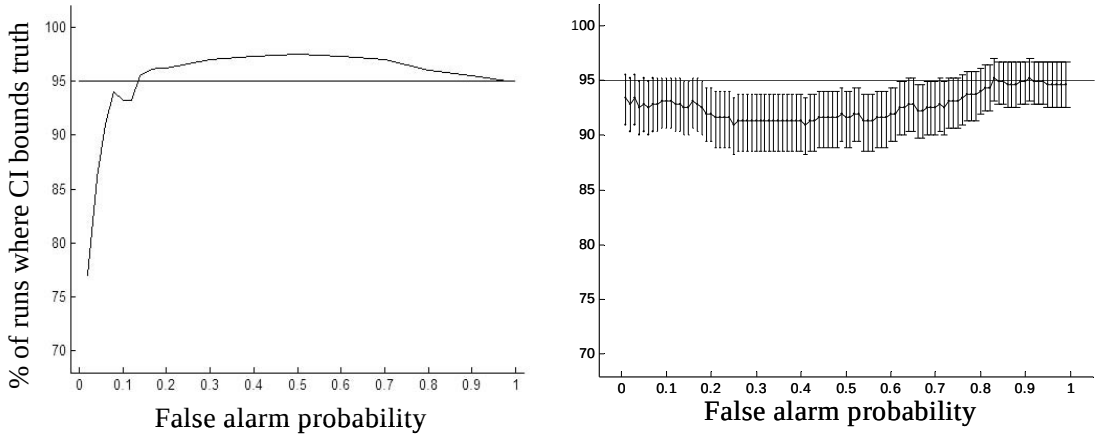


Figure 5.8 The ROC curve confidence interval coverage accuracies for the Hall [Hall *et al.*, 2004] method and the method developed here for beta target and non-target densities. Beta target and non-target densities generate 100 target samples and 100 non-target samples. This process is repeated many times to determine coverage accuracy. For the target density the beta parameters are a = 2 and $b = 4$, and for the non-target density they are $a = 2$ and $b = 3$. These figures otherwise use the same process as Figure 5.7. The left plot shows Hall's results and the right plot shows results of the method developed here.

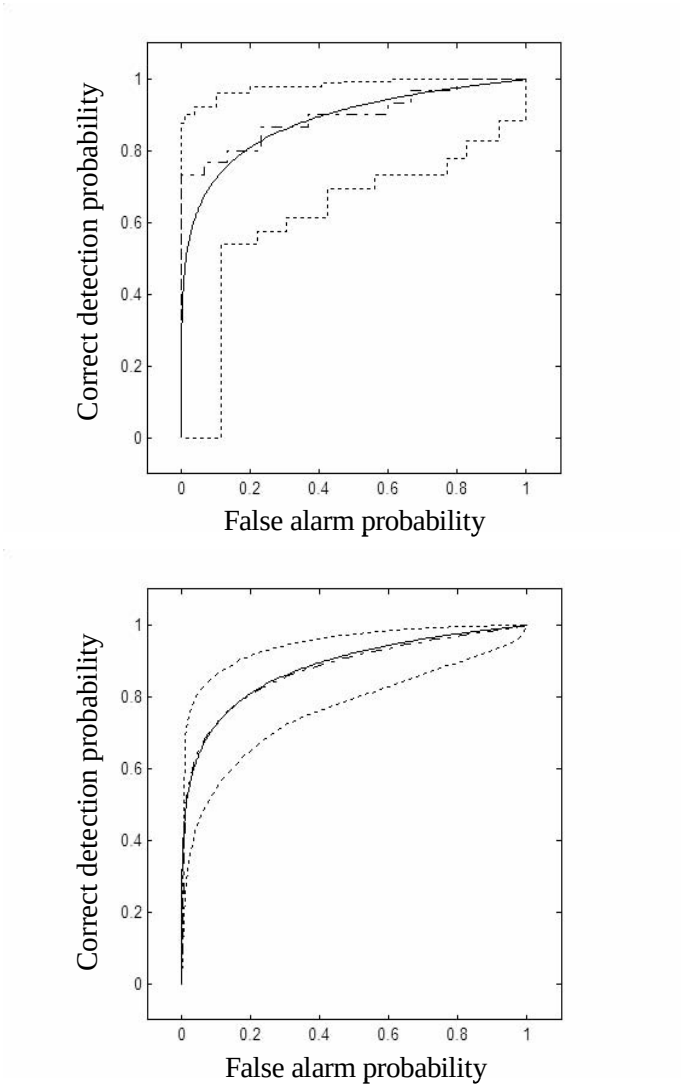


Figure 5.9 Comparison with the Hilgers [Hilgers, 1991] binomial method. The method uses techniques similar to the current AFRL approach for generating ROC curve confidence interval estimates. The top plot shows the 95% confidence intervals for the Hilgers method. These intervals cover the stated confidence interval region, but the confidence intervals are too wide [Schafer, 1994]. The bottom plot shows (also for 95% confidence intervals) that the approach developed here provides a smoother estimate of the ROC curve (dash/dotted line), and, more significantly, it produces much narrower confidence intervals.

conservative in that they are typically larger than necessary. Note that such a band is less informative than a band with smaller confidence band area provided that both bands have at least the stated coverage (95% in this case). The top plot of Figure 5.9 shows Hilgers' results for 30 target samples and 30 non-target samples obtained using Medcalc statistical software (commercially available software that implements Hilgers' approach in a 2005 update). The bottom plot shows a much narrower confidence band for the same samples obtained using the method developed here. In addition to the larger band width, the Hilgers approach also has a general disadvantage in that the rectangular region connection that forms the confidence band is generated by an ad-hoc method.

The results demonstrate the robustness of the method developed here when the overall model density form assumptions are correct. The method developed here is expected to improve ROC confidence interval results compared with other approaches in most cases. The method developed here provides a flexible and robust framework by which target and non-target samples, model assumptions, and prior densities can be incorporated.

5.5 Additional considerations

In determining which ROC confidence interval approach(es) are appropriate, sample size and knowledge of the density model form are important factors to consider. The following provides a few scenarios.

Large numbers of samples are available and there is no prior knowledge of target and non-target density form. Bootstrap methods may be acceptable. For example, the bootstrap method of Zhou [Zhou and Qin, 2005] may be acceptable, if a large number (more than 100) of target and non-target scores are available, and if the form of the target and non-target scores are not known, but are thought to be non-normal and non-beta. Figure 5.10 is similar to the Zhou method (Example 2/3) of Figure 5.5, except that rather than 20 target and 20 non-target samples, various numbers of samples are shown. Note that while the coverage accuracy improves for increased number of samples, a large

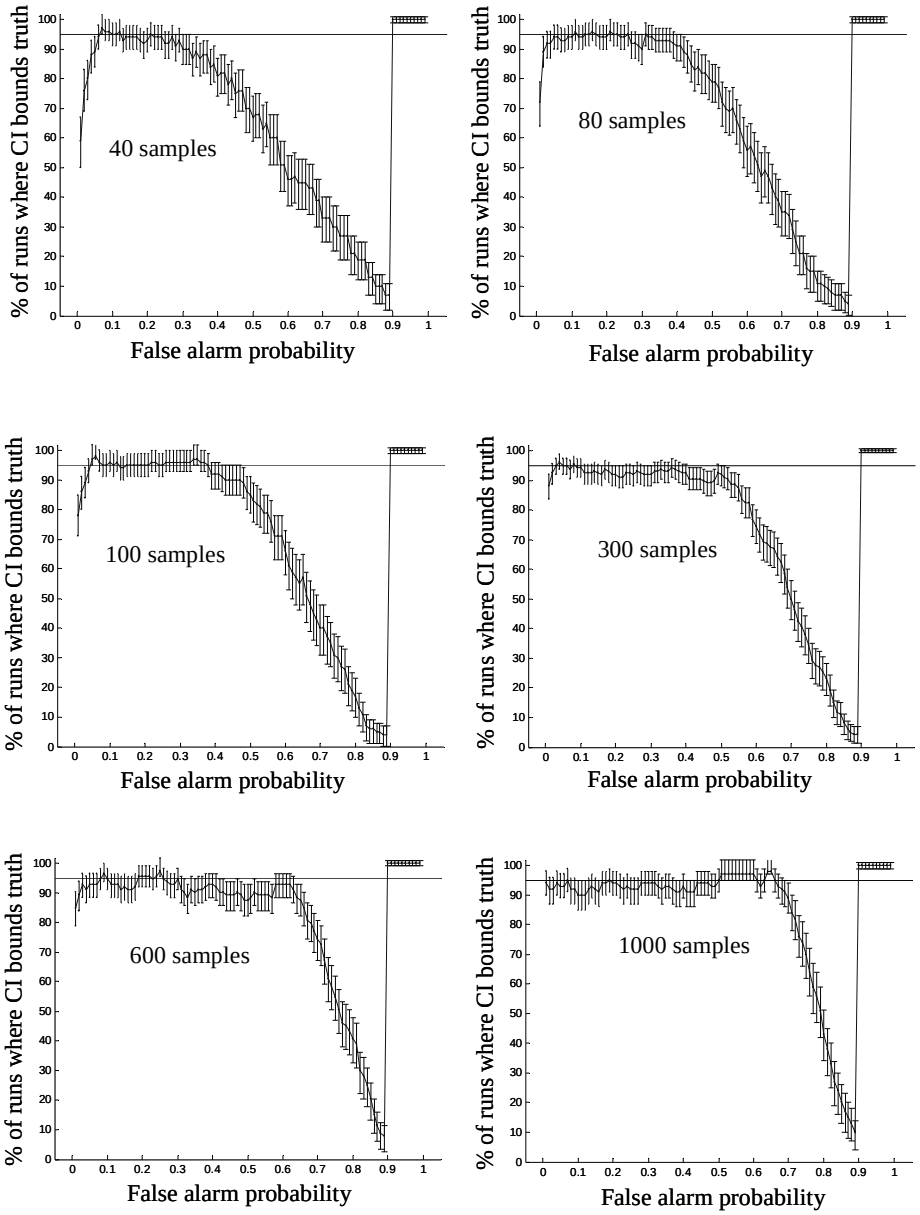


Figure 5.10 Coverage accuracy for Zhou confidence bounds for various numbers of target and non-target samples for a beta density model. The plots shown are the same as the bottom left plot of Figure 5.5, except that instead of 20 target and 20 non-target samples, the number of samples is increased to 40 target and 40 non-target samples, 80 target and 80 non-target samples, etc. Note that while the coverage accuracy does improve for increased number of samples, a large number of samples can be required for to achieve good coverage accuracy.

number of samples may be required to achieve good coverage accuracy. Coverage accuracy depends on the target and non-target density being evaluated. An additional example is shown in Figure 5.11 where the Zhou method forms confidence intervals

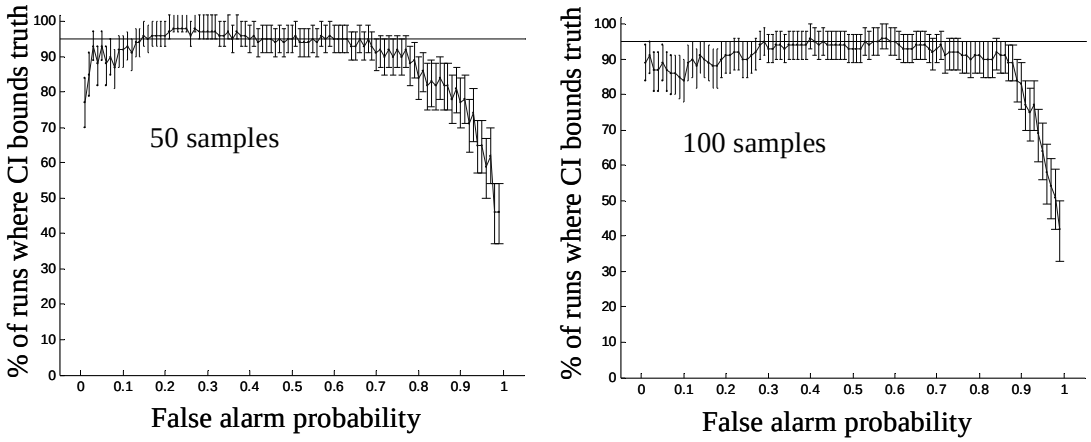


Figure 5.11 Coverage accuracy for Zhou confidence bounds for a normal density model. The plots shown use samples generated from the same underlying densities as Figure 5.7. Here runs of 50 target and 50 non-target samples and 100 target samples and 100 non-target samples are evaluated. The Zhou bootstrap method is used to obtain the displayed confidence intervals.

based on the samples generated from underlying normal densities (the same underlying densities previously used in Figure 5.7). For this example, the Zhou confidence bounds begin to provide appropriate coverage over most false alarm probabilities for somewhat lower numbers of samples. Thus, a paradox is introduced: the Zhou approach can provide appropriate coverage for "enough" samples, but in order to known how many samples are "enough" some knowledge of the underlying densities is needed.

Low numbers of samples are available, there is no prior knowledge of target and non-target density form, and highly conservative confidence bands are acceptable. Here the Hilgers [Hilgers, 1991] method is an appropriate choice.

Low numbers of samples are available, target and non-target densities are known to be normal or normal by some transformation, and the probability of target given score is

known to monotonically increase for increased score. The binormal approach, which attempts such assumptions (see Section 2.7.1), may be appropriate in this case.

The objective of the comparison detailed in the previous section is to demonstrate the viability of the framework developed here, not to prove that a selected model that uses this framework outperforms other approaches in every case (particularly when the selected model is not correct). Also, a key objective of the research here is to develop a performance metric uncertainty estimation approach that extends to the CEG curve.

The amount of time to execute a run (i.e. to move from a set of target and non-target samples to obtaining a confidence band) must also be considered. For the method developed here, two primary factors contribute to run time.

First, consider the computation of target and non-target posterior parameter densities, which are developed prior to any ROC curve formulation. The time to approximate posterior parameter densities depends on the number of target and non-target parameter points selected. Consider the parameter point selection process. For the beta density model, the process implemented here starts with 300 target points uniformly selected over mean and standard deviation and 300 non-target points also uniformly selected over mean and standard deviation. Then the combined posterior weightings are found for the sample values (see Equation (3.14)). The 16 grid point combinations that are closest to the mean and standard deviation of the samples are kept (4 target points, and 4 non-target points), along with any combinations that are greater in combined posterior weighting to any of these combinations. Then a 10 x 10 grid (100 points) for target means and standard deviations and a 10 x 10 grid (100 points) for non-target means and standard deviations is formed over this region, with much smaller grid point spacing. Again the combined posterior parameter weightings are found for each of the 10,000 grid point combinations, and only those points that contribute to 99.9% of the total posterior parameter weighting among these combinations are retained. The retained posterior parameter weightings then comprise an even smaller region than the previous iteration.

A second 10 x 10 grid (100 points) for target means and standard deviations and a 10 x 10 grid (100 points) for non-target means and standard deviations is then found. Again, the grid points that contribute to 99.9% of the total posterior parameter weighting among the 10,000 combinations of grid points are retained. The above operations for an example set of 20 target and 20 non-target samples takes approximately 70 seconds using the Matlab code developed here.

The second factor is ROC curve computation time. Each of the retained target and non-target grid point combinations form ROC curves, and these ROC curves must be computed (see Figure 3.8). Computation of each ROC curve takes approximately 0.75 seconds; the total run time for this section depends on the number of grid point combinations that make up the 99.9% of the final set of grid points (which can range from approximately 200 to 10000). Total run time for 20 target samples and 20 non-target samples generated from the densities of Zhou example 2/3 (see Figure 5.4) for a single example run is 244 seconds (assuming the beta density model process described here). Total run time for 50 target samples and 50 non-target samples for this same type of run is 251 seconds. In comparison, a method that implements Zhou's process (adjusted bootstrap with 250 bootstrap replications) in Matlab takes 15.5 seconds for the same 20 target and 20 non-target samples and 33.5 seconds for the same 50 target samples and 50 non-target samples. Samples generated from other density pairs can take significantly longer per run for the method developed here. A similar process, again for Zhou's example 2/3 for the CEG curve, takes 170 seconds for 20 target and 20 non-target samples and 154 seconds for 50 target and non-target samples. An increase in target and non-target samples can result in fewer grid point combinations in the final 99.9%, so run time may decrease with increase in samples. Also, the computation of a particular CEG curve (required for each grid point retained in the final set) is faster than the computation of a ROC curve, so the process is faster for CEG curve confidence intervals than ROC curve confidence intervals. An increase in number of target samples leads to a more highly peaked posterior probability density weighting, so the number of grid points used

may need adjustment as sample size changes. Figure 5.12 shows coverage examples; note that results converge as grid point spacing increases.

More complex density models (such as beta mixture models) can require significantly more grid points to cover the entire relevant parameter space (note also that the regions of high density weighting may be disjoint). Also, as the number of grid points becomes large, computation time increases proportional to the number of grid points squared; a small increase in number of grid points results in a large increase in run time.

Most of the computational challenges in terms of run time are apparent when attempts are made to verify results by determining coverage accuracy (e.g., the confidence band development process is repeated many times, such as 100 or more sets of 30 target and 30 non-target samples generated from the same underlying target and non-target densities).

Appendix C includes code to generate ROC curve and CEG curve confidence intervals. The appendix provides code for the beta density model, along with code for the two-beta mixture model. For the two-beta mixture models there are significantly more parameters (five versus two for the single beta model), and the above grid point iteration procedure is not applied. Instead, the process selects two-beta grid points at random, and calculates the combined posterior weighting for such grid points. The user specifies the number of random grid points for the two-beta mixture model; a typical number is 10000. The number of random grid points may be increased until convergence is observed. The number of points necessary for convergence depends on the specific sample values. Matlab matrix size limitations constrain the number of grid points to about 20000 (depending on the specific sample values). Methods available to improve run time are noted in the Future Work discussion.

Here the uncertainty estimation methods developed in Chapters 3 and 4 were compared with the current literature. The next chapter provides a summary of the results of the research and also identifies areas of interest for future work.

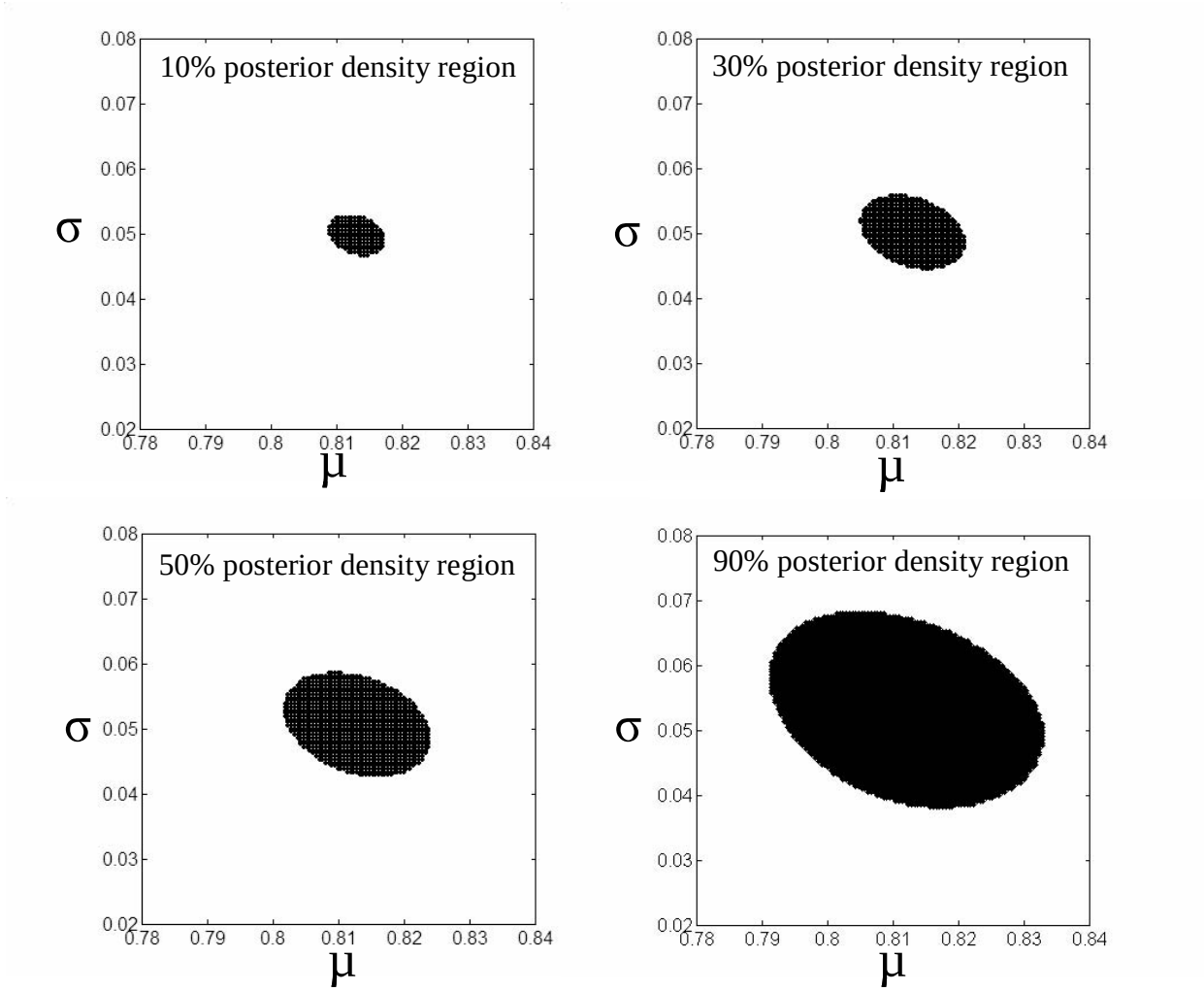


Figure 5.12 Regions that make up selected percentages of the posterior parameter density. The four plots show the regions that encompass 10%, 30%, 50%, and 90% of posterior parameter weighting for an example where a set of samples is generated from a target beta density.

6. Accomplishments, Contributions, and Future Work

Section 6.1 reviews the accomplishments and contributions of this research, and Section 6.2 describes areas of interest for future work.

Prior to listing the specific accomplishments of this research (in the next section), the results of the research presented here are placed in a proper perspective.

The primary contributions of this work are the framework described most fully in Chapter 3. Theorem 3.2, "ROC curve density", develops an analytical approach for forming the posterior probability density of the ROC curve. This theorem enables an exact description of the ROC curve probability density for given target and non-target samples, density model assumptions, and prior densities of model parameters. Theorem 3.3, "Numerical approximation of ROC curve density", extends this analytical description to a form that is computationally practical. Also important as a primary accomplishment is the extension of the probability density developments in Chapter 3 to confidence intervals (as described in Section 4.1.3).

The potential usefulness of the framework is further emphasized through a verification and evaluation process that includes comparisons with other methods. While the comparisons are interesting, it is improper to place undue emphasis on the results of the verification and evaluation process (Chapters 4 and 5) as primary contributions of this research, even though these results show promise. The theorems and further descriptions of Chapters 3 and 4 enable "actual probability density statements" [see Carlin, 2000, pp. 35-36] for a single set (or run) of target and non-target score samples, for given models, and for given prior assumptions.

Thus, there is no need to evaluate results based on the method developed here over many runs, although such runs can indicate efficacy. The exactness over one run of the Bayesian approach is arguably more importanty than what occurs "on average" over

many runs. Alternative methods of obtaining confidence intervals can be accurate (on average) over many runs, but make no claim regarding the results of any particular run. The approach introduced here enables an actual probability statement to be made from only one run, but it is not possible to verify or evaluate correctness except over many runs. Obviously, it would be desirable to have a process that provides an actual probability statement for one run and that also behaves appropriately over many runs (which the method developed here clearly does). In considering which approach is best, note that there will often be only one set of samples, so making the most appropriate statement possible based on only one run is arguably more important than what occurs over many runs.

The density model example assumed in this research is beta-based (predominantly focused on a unimodal beta density model). This model is merely as an example application of Theorems 3.2 and 3.3. Because the scores that are inputs in this research are continuous between zero and one, the beta density seems appropriate (see [Kagan *et al.*, 1973]); however, this research has not and does not intend to show that the beta density is effective and/or appropriate when the model density is not known. In particular, it is not the objective of this research to show that the beta density always provides a good estimate for all sets of data samples when model form is not known; the beta density model is simply an example. Thus, a caution on the results in Chapter 5 is that the comparisons with existing research do not enable true "apples-to-apples" comparisons; the comparisons made in Chapter 5, while appropriate in demonstrating the Bayesian framework, do not show that the method developed here is necessarily an improvement over existing approaches. In simply demonstrating the framework, the method developed here generally uses samples from beta densities where the parameters are assumed to be unknown. (The method developed here then uses the samples to develop probability densities for the unknown parameters.) Also, the comparisons are generally made with methods that make differing model assumptions. Note that currently available methods in the literature do not enable the selection of a beta density

model. As discussed below and in future work, it would be of interest for future developments to incorporate the results of Chapters 3 and 4 into flexible software that enables user selection of densities or model assumptions.

Unless one is guaranteed that a particular model assumption or prior is correct, a reasonable question concerns the usefulness of the results of the framework developed here. Consider the available alternatives. Bootstrap based approaches avoid assumptions, but as is shown in Figure 5.10, unless large numbers of samples are available, avoiding such assumptions can yield poor results (certainly if large numbers of samples are available, then bootstrapping based approaches are very much of interest). Existing research, with the exception of bootstrap-based approaches, make model assumptions; the framework presented here also makes model assumptions. The difference between the method developed here and other approaches is that the other approaches develop frameworks that involve restrictive model assumptions. The framework developed here enables flexible model assumptions. In this regard (as future work) the framework developed here could be extended so that, for example, the user might specify "bi-modal density mixture model", "tri-modal beta density mixture model", etc.

Another question is that if it is not known whether or not a set of samples is modeled well by a beta density model, how could the research presented here possibly be of interest? Two considerations are as follows. First, as future work, an examination of the fit of a beta density model to experimental data with fixed end points is of interest. Second, an extension that also may be of interest for future work is the incorporation of models of varying complexity, which is possible through regularization (see [Bishop, 1995]) and the use of the Occam factor (see [Gregory, 2005]). Such approaches do not select a beta density model or a bi-modal density model, etc., instead they incorporate models of different complexities; less complex models, such as single beta densities, receive higher overall weighting, more complex models receive less overall weighting (even though

there may be specific instances where such weightings fit the data better). The use of roughness is an alternative approach that incorporates models of various complexity (see discussion in Future Work, and related results presented in Appendix B).

6.1 Accomplishments and contributions

This research applies a new framework for ROC curve uncertainty estimation that is fully Bayesian, that is numerically tractable, and that leads to substantial improvements over existing methods. Quantitative comparisons are made; however, qualitative improvements are the most important outcome of the research presented here. As discussed in Chapters 2 and 5, most existing methods make restrictive assumptions that inhibit the application of a flexible model framework as presented here; the bootstrap approaches do not require such assumptions but are of limited applicability for small numbers of samples.

A significant aspect of this research is that the uncertainty estimation process developed here transitions to CEG curves. The CEG curve is a critical metric for AFRL in determining the usefulness of target detection systems. With a typically limited amount of data and with no appropriate methods for CEG curve uncertainty estimation, AFRL has previously been able to make only limited use of this metric. With the methods developed here, the CEG curve can be applied and its uncertainty can be estimated even for low numbers of samples.

The research reported here demonstrates the application of ROC curve uncertainty estimation methods from the medical community to target detection. It also provides more comprehensive qualitative and quantitative comparisons of alternative ROC curve and AUC value uncertainty estimation approaches than any available in the literature.

ROC curve density and confidence interval generation. This research applies a Bayesian framework to develop new methods for ROC curve density generation which

are also applicable to other target detection performance metrics. The framework is provided within Chapter 3 (which includes four theorems, a lemma and a procedure); more specifically, Theorems 3.2 provides an analytical approach for forming the probability density of the ROC curve, and Theorem 3.3 extends this analytical description into a form that is practical to evaluate analytically. Note that while ROC curve definitions are examined in the previous literature (see [Lloyd, 2002] and [Zhou and Qin, 2005]), the probability density results obtained here are unprecedented.

Computations of confidence bands or confidence intervals (as described in Section 4.1.3) can be made from the performance metric densities in a straightforward manner. This capability contrasts with previous methods in the literature, which generally are applicable only to specific band or interval definitions and which can not be easily extended. Application of the Bayesian framework allows the user of a SUT to better understand conclusions from performance metrics, especially if they are based on limited data.

This research presents the results of simulations and real-data experiments that demonstrate the significance of the new uncertainty estimation methods. Computational techniques that implement the methods are demonstrated, and they are shown to yield accurate results that are otherwise not analytically tractable. Significantly, the methods developed here enable the calculation of actual performance metric probability densities for given target and non-target score samples, given density forms for the scores, and given prior densities for the parameters in these forms.

Representative ROC curve generation. This research develops methods that generate representative ROC curves (samples from a ROC curve density) from given sets of target and non-target samples. Numerical implementation of the method for generating the ROC (and CEG) curve densities results in the generation of representative ROC (and CEG) curves. Macskassy [Macskassy and Provost, 2004] [Macskassy *et al.*, 2005] most

recently emphasizes the critical need for such representative ROC curves and the lack of such ROC curves in the literature. From such representative ROC curves (or representative CEG curves), many descriptive statistics, such as mean and median ROC curves and AUC values and confidence bands and intervals for them, are obtained. The results are shown to be robust when the overall model density form assumptions are correct.

CEG curve density, representative CEG curve generation, and confidence interval generation. The methods developed here can be applied to CEG curves. The lack of a proven means for obtaining confidence intervals for the CEG curve was a primary motivation for AFRL sponsorship of this research. The research reported here goes beyond simply adapting an existing ROC curve confidence interval estimation method and applying it to the ROC curve. Instead, it applies a Bayesian framework to create, demonstrate, and validate new methods that can be applied beyond the uncertainty estimation problem originally addressed.

Target and non-target density flexibility. Although the examples considered here use beta densities, the methods developed here can be directly applied to other density forms. In contrast, the binormal ROC curve in predominant use implies a nearly convex ROC curve form and restricts curve estimation to this form. The methods developed here are particularly important for cases where sample size is small, as is typical in target detection problems. Thus, this research is expected to alter the way that the target detection evaluation community approaches ROC and CEG curve uncertainty estimation.

6.2 Future work

The success of this research should motivate further investigation in several areas:

1. *Improve the efficiency of target and non-target density posterior parameter computation.* As the number of parameter evaluation points increases (see Figure 4.9), the ROC curve density converges (see Theorem 3.3) provided that the relative spacing of the points does not change (for example, the spacing is kept uniform over mean and standard deviation). More computationally efficient methods to obtain sufficient numbers of evaluation points should be investigated. This optimization would assist in the transfer of the Bayesian framework to more complex density models. Jordan [Jordan *et al.*, 1999] focuses on a variational approach and references alternatives such as the pruning algorithm, bounding conditioning, search-based methods, and localized partial evaluation. Bos [Bos, 2002] describes alternatives such as Gibbs sampling and importance sampling. Madigan [Madigan and Raftery, 1994], Raftery [Raftery *et al.*, 2003], and Hoeting [Hoeting *et al.*, 1999] reference Bayesian model averaging and Occam's window for reducing the computational complexity of posterior parameter density evaluation.

2. *Develop integrated confidence band computation approaches.* As noted in Section 2.7, while the framework used and the methods developed here apply to many types of ROC curve uncertainty estimation, there are other approaches that may be acceptable in particular cases. For example, the binomial approach provides bands that encompass greater than or equal to 95% coverage for 95% confidence bands. Confidence bands based on the binomial approach are overly conservative but may be applied as an upper bound to ROC curve confidence bands for the method developed here. Thus, relevant aspects of each of the approaches may be combined to achieve joint-method ROC curve confidence bands.

3. *Test the methods developed here with other density models.* Example alternative density models include hybrid models that combine Gaussian densities, beta densities, or both. In such a combination approach, density models that have higher complexity, even if they fit the data well, may be regarded as less likely to represent the true model (see

[MacKay, 1992b]). Here complexity could refer to the number of parameters in the model, e.g., a single-beta density has two parameters (mean and standard deviation) and a two-beta mixture density has five parameters (two means and standard deviations plus an amplitude ratio). Regularization techniques can combine models of varying numbers of parameters (see Bishop [Bishop, 1995]). To avoid the possible over-fitting effects of more complex densities, target and non-target score density function roughness or ROC or CEG curve roughness could be used to quantify complexity. Appendix B addresses related issues by first examining interpolation methods that have desirable extrapolation properties based on roughness; it then describes an analytical approach for roughness computation, where the roughness of a function is defined as its integrated squared second derivative. Approaches that incorporate roughness recognize, for example, that a density function with large roughness that describes the data well may be less desirable than a density function that describes the data less well but that has low roughness.

4. *Apply the methods developed here to additional performance metrics.* Once the ROC curve density is developed, the research presented here shows that transition to the CEG curve density is straightforward. This transition could be made to other performance metrics, including the Dice similarity coefficient (see Zou [Zou *et al.*, 2004]), mutual information (see Zou [Zou *et al.*, 2004]), partial AUC (see [Dodd and Pepe, 2003]), and the Youden index (see Faraggi [Faraggi, 2003]).

Appendix A. Analytical Derivations and Numerical Approximations

A.1 Derivation of ROC curve

Theorem Score-threshold ROC curve

Let $f(s;u)$ and $g(s;v)$ be densities of s given u and v, where s is a scalar and u and v are vectors, assume that $f(s;u)$ and $g(s;v)$ are integrable $\forall\ u, v$ and let $\widehat{F}(t;u) = \int_t^\infty f(s;u)ds$ and $\widehat{G}(t;v) = \int_t^\infty g(s;v)ds$. Also let $w = [u_1\ u_2\ ...\ v_1\ v_2\ ...]$ and $\widehat{F}(t;u) = 1 - F(t;u)$ and $\widehat{G}(t;u) = 1 - F(t;u)$ where $F(t;u)$ and $G(t;v)$ are cumulative probability distributions. Let $x = \widehat{F}(t;u)$ and $y = \widehat{G}(t;v)$. Assume there is a unique correspondence of s to $\widehat{F}(s;u)$ such that $0 \leq \widehat{F}(s;u) \leq 1$ and $\widehat{F}^{-1}$ is invertible (by the Implicit and Inverse function theorems; see [Olmstead, 1961]). Then $y = r(x;w)$, where $r = \widehat{G}\widehat{F}^{-1}$.

Proof

If $F(s;u) \equiv \int_{-\infty}^{\widetilde{s}} f(\widetilde{s};u)ds$ is a cumulative distribution function (CDF), then [Stark and Woods, 1986, pp. 42]

$$P_u(s_1 < X < s_2) = F(s_2;u) - F(s_1;u)\ \geq 0 \text{ for } s_1 < s_2. \tag{A.1}$$

If $f(s;u)$ is a probability density function (PDF), then [Stark and Woods, 1986, pp. 44]

$$\int_{s_1}^{s_2} f(s;u)ds = P_u(s_1 \leq S \leq s_2). \tag{A.2}$$

By Equations (A.1) and (A.2),

$$\int_{s_1}^{s_2} f(s;u)ds = F(s_2;u) - F(s_1;u). \tag{A.3}$$

Also [Stark and Woods, 1986, pp. 41],

$$F(-\infty; u) = 0,\ F(\infty; u) = 1. \tag{A.4}$$

By Equations (A.3) and (A.4),

$$\int_{s_i}^{\infty} f(s; u)ds = F(\infty; u) - F(s_i; u) = 1 - F(s_i; u). \tag{A.5}$$

Since it has been defined that $x = \int_t^{\infty} f(s; u)ds = \widehat{F}(t; u)$, by Equation (A.5)

$$x = \int_t^{\infty} f(s; u)ds = 1 - F(t; u) = \widehat{F}(t; u). \tag{A.6}$$

Using an identical argument,

$$y = \int_t^{\infty} g(s; v)ds = 1 - G(t; v) = \widehat{G}(t; v). \tag{A.7}$$

Further, since $F(s; u) = 1 - \widehat{F}(s; u)$, and since $F(s_1; u) \leq F(s_2; u)$ for $s_1 \leq s_2$,

$$\widehat{F}(s_1; u) \geq \widehat{F}(s_2; u). \tag{A.8}$$

Since F is continuous from the right [Stark and Woods, 1986, pp. 44], i.e.,
$F(s; u) = \lim_{\epsilon \to 0} F(s + \epsilon; u), \epsilon > 0$, and since $F(s; u) = 1 - \widehat{F}(s; u)$,

$$\widehat{F}(-\infty; u) = 0;\ \widehat{F}(\infty; u) = 1, \tag{A.9}$$

$\widehat{F}$ is continuous from the left, i.e., $\widehat{F}(s; u) = \lim_{\epsilon \to 0} \widehat{F}(s - \epsilon; u), \epsilon > 0$.

Since $\widehat{F}^{-1}$ is invertible and unique for each u, then for any x, $0 \leq x \leq 1$, $x = \int_t^\infty f(s;u)ds$ for some unique threshold $t = \widehat{F}^{-1}(x;u)$, $\widehat{G} \circ \widehat{F}^{-1}(x;w) = \widehat{G}(\widehat{t};u)$, $x = \widehat{F}(t;u)$, and it follows that $y = r(x;w)$, where $r = \widehat{G} \circ \widehat{F}^{-1}(x;w)$.

Comments

If $f(s;u)$ and $g(s;v)$ are modeled by beta probability densities, then u and v are two-element vectors, and

$$f(s;u) = \frac{s^{\widetilde{u}_1-1}(1-s)^{\widetilde{u}_2-1}}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1+\widetilde{u}_2)}},\ 0 \leq s \leq 1, \tag{A.10}$$

$$g(s;v) = \frac{s^{\widetilde{v}_1-1}(1-s)^{\widetilde{v}_2-1}}{\frac{\Gamma(\widetilde{v}_1)\Gamma(\widetilde{v}_2)}{\Gamma(\widetilde{v}_1+\widetilde{v}_2)}},\ 0 \leq s \leq 1, \tag{A.11}$$

where $\widetilde{u}$ and $\widetilde{v}$ are related to u and v by

$$\widetilde{u}_1 = u_1[\frac{u_1(1-u_1)}{u_2} - 1] \tag{A.12}$$

$$\widetilde{u}_2 = \widetilde{u}_1[\frac{1}{u_1} - 1] \tag{A.13}$$

$$\widetilde{v}_1 = v_1[\frac{v_1(1-v_1)}{v_2} - 1] \tag{A.14}$$

$$\widetilde{v}_2 = \widetilde{v}_1[\frac{1}{v_1} - 1] \tag{A.15}$$

and where

$$\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt, a > 0. \tag{A.16}$$

Thus, $x = \widehat{F}(t;u) = \int_t^\infty f(s;u)ds$ may be expressed

$$x = \int_t^1 \frac{s^{\widetilde{u}_1-1}(1-s)^{\widetilde{u}_2-1}}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1+\widetilde{u}_2)}}ds, \tag{A.17}$$

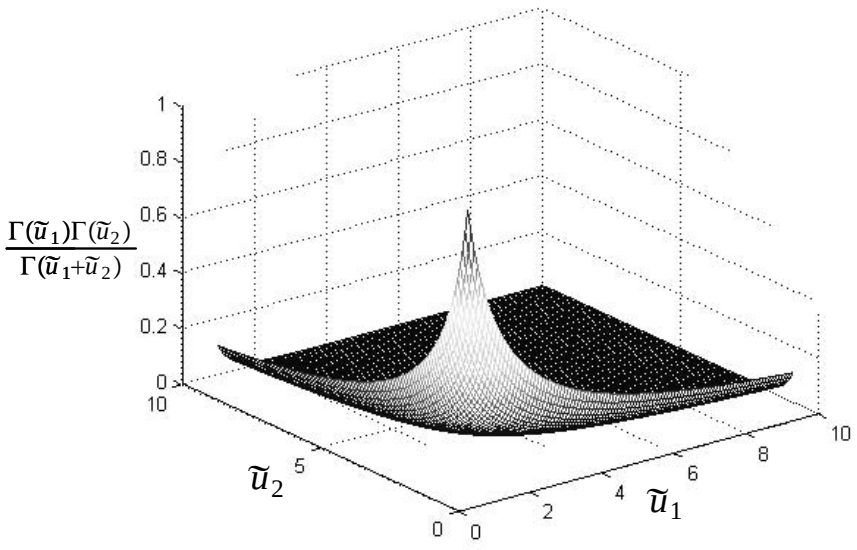


Figure A.1 Here $\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1+\widetilde{u}_2)}$ is shown as a function of $\widetilde{u}_1$ and $\widetilde{u}_2$.

where t is a selected threshold and $0 \leq t \leq 1$. Figure A.1 shows the relation of $\widetilde{u}_1$, $\widetilde{u}_2$, to $\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1+\widetilde{u}_2)}$.

Evaluation using Weierstrass' product [Korn and Korn, 2000, pp. 822], shows that $\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ may be factored into the infinite sum

$$\frac{1}{\Gamma(z)} = ze^{Cz}\prod_{k=1}^{\infty}[(1+\frac{z}{k})e^{-z/k}], \tag{A.18}$$

where C ≈ 0.5772157 is the Euler-Mascheroni constant and

$$\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \frac{a+b}{ab}\prod_{k=1}^{\infty}\frac{(k+a+b)(k)}{(k+a)(k+b)}. \tag{A.19}$$

Note (see [Patel *et al.*, 1976]) that

$$\int_0^t \frac{s^{\widetilde{u}_1-1}(1-s)^{\widetilde{u}_2-1}}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1+\widetilde{u}_2)}}ds = \frac{B_t(\widetilde{u}_1,\widetilde{u}_2)}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1+\widetilde{u}_2)}} = I_t(\widetilde{u}_1,\widetilde{u}_2), \tag{A.20}$$

where

$$B_t(\widetilde{u}_1, \widetilde{u}_2) = \int_0^t s^{\widetilde{u}_1 - 1}(1-s)^{\widetilde{u}_2 - 1} ds \tag{A.21}$$

and I_t is the incomplete beta function ratio. Also note that

$$\int_t^1 \frac{s^{\widetilde{u}_1 - 1}(1-s)^{\widetilde{u}_2 - 1}}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1 + \widetilde{u}_2)}} ds = 1 - \int_0^t \frac{s^{\widetilde{u}_1 - 1}(1-s)^{\widetilde{u}_2 - 1}}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1 + \widetilde{u}_2)}} ds \tag{A.22}$$

so that

$$\int_t^1 \frac{s^{\widetilde{u}_1 - 1}(1-s)^{\widetilde{u}_2 - 1}}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1 + \widetilde{u}_2)}} ds = 1 - I_t(\widetilde{u}_{1,}\widetilde{u}_2). \tag{A.23}$$

For the incomplete beta function ratio [Patel *et al*., 1976, pp. 246]

$$I_t(\widetilde{u}_{1,}\widetilde{u}_2) = 1 - I_{1-t}(\widetilde{u}_{2,}\widetilde{u}_1), \tag{A.24}$$

so that

$$\int_t^1 \frac{s^{\widetilde{u}_1 - 1}(1-s)^{\widetilde{u}_2 - 1}}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1 + \widetilde{u}_2)}} ds = 1 - (1 - I_{1-t}(\widetilde{u}_{2,}\widetilde{u}_1)). \tag{A.25}$$

Therefore

$$\int_t^1 \frac{s^{\widetilde{u}_1 - 1}(1-s)^{\widetilde{u}_2 - 1}}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1 + \widetilde{u}_2)}} ds = I_{1-t}(\widetilde{u}_{2,}\widetilde{u}_1). \tag{A.26}$$

From above, and noting that Equations (A.12)-(A.14) may be manipulated to solve for u_1, $u_{2,}v_{1,}$ and v_2 using $u_1 = \frac{\widetilde{u}_1}{\widetilde{u}_1 + \widetilde{u}_2}$, $u_2 = \frac{\widetilde{u}_1\widetilde{u}_2}{(\widetilde{u}_1+\widetilde{u}_2+1)(\widetilde{u}_1+\widetilde{u}_2)^2}$, $v_1 = \frac{\widetilde{v}_1}{\widetilde{v}_1+\widetilde{v}_2}$, and $v_2 = \frac{\widetilde{v}_1\widetilde{v}_2}{(\widetilde{v}_1+\widetilde{v}_2+1)(\widetilde{v}_1+\widetilde{v}_2)^2}$,

$$x = I_{1-t}(\widetilde{u}_{2,}\widetilde{u}_1) = \widehat{F}(t; \frac{\widetilde{u}_1}{\widetilde{u}_1 + \widetilde{u}_2}, \frac{\widetilde{u}_1\widetilde{u}_2}{(\widetilde{u}_1 + \widetilde{u}_2 + 1)(\widetilde{u}_1 + \widetilde{u}_2)^2}), \tag{A.27}$$

and similarly

$$y = I_{1-t}(\widetilde{v}_{2,}\widetilde{v}_1) = \widehat{G}(t; \frac{\widetilde{v}_1}{\widetilde{v}_1 + \widetilde{v}_2}, \frac{\widetilde{v}_1\widetilde{v}_2}{(\widetilde{v}_1 + \widetilde{v}_2 + 1)(\widetilde{v}_1 + \widetilde{v}_2)^2}). \tag{A.28}$$

Therefore, for a beta density model and for given values of u and v,

$$y = r(x; w) = \widehat{G} \circ \widehat{F}^{-1}(x; w) = I_{1-\widehat{F}^{-1}(x;w)}(\widetilde{v}_2, \widetilde{v}_1) \text{ and } \widehat{F}(t; u) = I_{1-t}(\widetilde{u}_2, \widetilde{u}_1). \quad \text{(A.29)}$$

Thus, whereas there are various ways to describe r (such as an infinite series of products, gamma functions, and the incomplete beta function ratio), such expressions are impractical to further evaluate analytically. Even if they were practical to evaluate, the analytical expressions would be for the ROC curve, not for the ROC curve density.

A.2 Derivation of ROC curve density

Theorem 3.2 ROC curve density

Let $d = \{s_i : i = 1, ..., I\}$ be a set of independent and identically distributed samples s_i from distribution f. Let $h = \{q_j : j = 1, ..., J\}$ be a set of independent and identically distributed samples q_j from distribution g, and let $p_u(u)$ and $p_v(v)$ be prior densities of the random parameter vectors u and v. Let $\widetilde{A}$ be the admissible set of u and v parameters. Then

$$p_{y|x}(y|x, d, h) = \widetilde{C}_0 \iint_{\widetilde{A}} p_{y|x}(y|x, u, v) \prod_i f(s_i; u) \prod_j g(q_j; v) p_u(u) p_v(v) du dv, \quad \text{(A.30)}$$

where the constant $\widetilde{C}_0$ depends on d and h.

Proof

Let w be the concatenation of u and v (i.e., $w = [u_1\ u_2\ ...\ v_1\ v_2\ ...]$), and let D be the concatenation of d and h.

By marginalization

$$p_{y|x}(y|x, D) = \int_{\widetilde{A}} p_{y|x}(y|x, w) p_{wD}(w|D) dw, \quad \text{(A.31)}$$

and by Bayes' rule,

$$p_{w|D}(w|D) = \widetilde{C}_1 p_{D|w}(D|w) p_w(w), \quad \text{(A.32)}$$

where the constant $\widetilde{C}_1$ depends on D.

Thus, by independence

$$p_{D|w}(D|w) = \widetilde{C}_2 \prod_{i=1}^{I} f(s_i; u) \prod_{i=1}^{J} g(q_j; v), \tag{A.33}$$

where the constant $\widetilde{C}_2$ depends on d and h.

and

$$p_w(w) = p_u(u)p_v(v). \tag{A.34}$$

Combining Equations (A.32), (A.33), and (A.34) shows that Equation (A.31) is equivalent to

$$p_{y|x}(y|x, D) = \widetilde{C}_0 \iint_{\mathcal{A}} p_{y|x}(y|x, u, v) \prod_i f(s_i; u) \prod_j g(q_j; v) p_u(u) p_v(v) \mathrm{d}u \, \mathrm{d}v, \tag{A.35}$$

where the constant $\widetilde{C}_0$ depends on d and h.

Note that $\widetilde{A}$ is used here rather than A because notation earlier in this document (see Equation (3.3)) refers to the admissible set for the beta density model as A, and this proof is not restricted to the beta density model.

Comments

For a beta density,

$$p_{yx}(y|x, D) = \widetilde{C}_0 \int_1^\infty \int_1^\infty \int_1^\infty \int_1^\infty p_{yx}(y|x, \widetilde{u}_1, \widetilde{u}_2, \widetilde{v}_1, \widetilde{v}_2) \prod_i f(s_i; \widetilde{u}_1, \widetilde{u}_2) \prod_j g(q_j; \widetilde{v}_1, \widetilde{v}_2)$$

$$\cdot p_u\left(\frac{\widetilde{u}_1}{\widetilde{u}_1+\widetilde{u}_2}, \frac{\widetilde{u}_1\widetilde{u}_2}{(\widetilde{u}_1+\widetilde{u}_2+1)(\widetilde{u}_1+\widetilde{u}_2)^2}\right) p_v\left(\frac{\widetilde{v}_1}{\widetilde{v}_1+\widetilde{v}_2}, \frac{\widetilde{v}_1\widetilde{v}_2}{(\widetilde{v}_1+\widetilde{v}_2+1)(\widetilde{v}_1+\widetilde{v}_2)^2}\right) d\widetilde{u}_1 d\widetilde{u}_2 d\widetilde{v}_1 d\widetilde{v}_2,$$

and $p_{yx}(y|x, \widetilde{u}_1, \widetilde{u}_2, \widetilde{v}_1, \widetilde{v}_2) = \delta(y - r(x; w))$,

where

$$r(x;\widetilde{v}_{2,}\widetilde{v}_1,\widetilde{u}_{2,}\widetilde{u}_1)=\widehat{G}\circ\widehat{F}^{-1}(x;\widetilde{v}_{2,}\widetilde{v}_1,\widetilde{u}_{2,}\widetilde{u}_1)=I_{1-\widehat{F}^{-1}}(\widetilde{v}_{2,}\widetilde{v}_1),\ \widehat{F}(t;\widetilde{u})=I_{1-t}(\widetilde{u}_{2,}\widetilde{u}_1) \tag{A.36}$$

and u and v are related to $\widetilde{u}$ and $\widetilde{v}$ by Equations (A.12)-(A.14). Also,

$$\prod_i f(s_i|\widetilde{u}_{1,}\widetilde{u}_2)=\prod_i \frac{s_i^{\widetilde{u}_1-1}(1-s_i)^{\widetilde{u}_2-1}}{\frac{\Gamma(\widetilde{u}_1)\Gamma(\widetilde{u}_2)}{\Gamma(\widetilde{u}_1+\widetilde{u}_2)}}$$

$$\prod_j g(q_j|\widetilde{v}_{1,}\widetilde{v}_2)=\prod_j \frac{q_j^{\widetilde{v}_1-1}(1-\widetilde{v}_i)^{\widetilde{v}_2-1}}{\frac{\Gamma(\widetilde{v}_1)\Gamma(\widetilde{v}_2)}{\Gamma(\widetilde{v}_1+\widetilde{v}_2)}}.$$

One example choice of parameter density prior has $p_u(u_{1,}u_2)$ equal to a constant over all values of u_1 and u_2 for which the beta density is defined, where u_1 is mean and u_2 is standard deviation. With an identical choice of priors for $p_v(v_{1,}v_2)$, the following bounds apply:

$p_u(u_{1,}u_2)=1, 0\le u_1\le 0.5$, and $u_2\le\frac{1-u_1}{u_1(u_1+2)(u_1+1)^2}$

$p_u(u_{1,}u_2)=1, 0.5\le u_1\le 1$, and $u_2\le\frac{u_1(1-u_1)^2}{2-u_1}$

$p_u(u_{1,}u_2)=0, 0\le u_1\le 0.5$, and $u_2>\frac{1-u_1}{u_1(u_1+2)(u_1+1)^2}$

$p_u(u_{1,}u_2)=0, 0.5\le u_1\le 1$, and $u_2>\frac{u_1(1-u_1)^2}{2-u_1}$

$p_v(v_{1,}v_2)=1, 0\le v_1\le 0.5$, and $v_2\le\frac{1-v_1}{v_1(v_1+2)(v_1+1)^2}$

$p_v(v_{1,}v_2)=1, 0.5\le v_1\le 1$, and $v_2\le\frac{v_1(1-v_1)^2}{2-v_1}$

$p_v(v_{1,}v_2)=0, 0\le v_1\le 0.5$, and $v_2>\frac{1-v_1}{v_1(v_1+2)(v_1+1)^2}$

$p_v(v_{1,}v_2)=0, 0.5\le v_1\le 1$, and $v_2>\frac{v_1(1-v_1)^2}{2-v_1}$.

Even for the case of uniform prior density over admissible mean and standard deviations and with single beta densities (simple in comparison with beta mixture models), there are

no less than six incomplete gamma functions inside the integral. Without considering the definition for $p_{y|x}(y|x, D)$, since the gamma function is itself analytically described by an integral, an analytical solution, even for a single beta density, is not feasible (multiple analytic terms inside the four part integral would consist of $\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt, a > 0$). However, using Monte Carlo methods, a convergent numerical result may be obtained. Further, rather than the restrictive solution that an analytical development would produce (restricted to single beta models), the numerical development may be extended to beta mixture models or other families of density models. Thus, based on the analytic framework it is clear that a numerical evaluation is needed. The evaluation points of Figure 3.6, shown within the allowed standard deviation versus mean plots, are sampling points used to estimate the full Bayesian posterior, which may be visualized as a three-dimensional density. The oval regions of the two left plots of this figure, shown in the vicinity of the target and non-target mean and standard deviation, indicate confidence interval bounds for the posterior probability. Similarly, the darkened regions of Figure 5.12 indicate 10%, 30%, 50%, and 90% confidence interval bounds for the posterior probability.

Appendix B. Analytical derivation of Roughness for Cardinal Interpolation

B.1 Introduction and background on cardinal interpolation

Gustafson, Parker, and Martin [Gustafson *et al.*, 2006] apply Bayesian methods to find the probability density of certain interpolating functions, where this density has desirable extrapolation properties that define cardinal interpolation. In this appendix, cardinal interpolation and roughness are introduced and then an analytical extension to Gustafson, Parker, and Martin [Gustafson *et al.*, 2006] is provided. As described in future work (Section 6.2), incorporating roughness into a target or non-target density model can provide a means to characterize and control models of various complexity for performance metric uncertainty.

Development of the cardinal interpolation density provided an early example for the development of densities for ROC and CEG curves that is the key advance reported here. Calculation of the cardinal interpolation density is facilitated by an analytical derivation of roughness of a sum of Gaussian functions, where roughness is defined as integrated squared second derivative of the sum of the functions. The use of roughness here is the degree of smoothness in Bishop (see [Bishop, 1995, pp. 173]). See [Bishop, 1995] and [MacKay 1992a, 1992b] for the related discussion of regularization.

The following summarizes the cardinal interpolation concept and its use of the analytical derivation of roughness. Gustafson, Parker, and Martin [Gustafson *et al.*, 2006] provide a full description.

The cardinal interpolation density combines a linear model with a Gaussian radial basis function model. When estimating points that are far from observed data points, an appropriate model is assumed to be a least squares line; when estimating points that are

close to observed data points, an appropriate model is assumed to be an interpolator (in this case a Gaussian radial basis function interpolator). Let data points D = $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ with $x_1 < x_2 < \ldots < x_n$ be samples from a Gaussian probability density in y relative to a line (see [Bishop, 1995]). By marginalization, the probability density $p(y|x, D)$ of y given x and D for a linear model is $\int p(y|x, a, b)p(a, b|D)da\ db$, where a is the intercept and b is the slope of the line. By Bayes' rule, $p(a, b|D)$ is proportional to $p(D|a, b)p(a)p(b)$ for independent a and b, where $p(D|a, b)$ is the product of $p(y|x, D)$evaluated at each of the data points and is thus proportional to the deviation weight $exp[-\sum(y_i - a - bx_i)^2/(2\sigma^2)]$. The result is a density for the linear model (see [Bishop, 1995]) that has a mean which is the least squares line at D.

The cardinal interpolation density uses the above linear model with a Gaussian radial basis interpolating model. The combined model is $y(x; a, b, c) = a + bx + \sum A_i exp[-(x - x_i)^2/(2c^2)]$, where each basis function has its mean at a point x value, has variance c^2, and has amplitude A_i such that $y_i = y(x_i; a, b, c)$ so that the points are interpolated. Regularization (see [Bishop, 1995]) yields weighting that depends on roughness $r(a, b, c)$. The cardinal interpolation density is developed by requiring that the roughness weight $exp(-Kr(a, b, c)$ equal the above deviation weight, where K is such that both types of weights have the same minimum.

B.2 Analytical roughness expression

The following expression for roughness has been verified for many sum of Gaussian functions using numerical integration. The use of this expression can greatly reduce the number of required computations as compared with numerical integration.

Theorem

Let $a, b \in \mathbb{R}$ and $c > 0$ and $x_{1,}x_{2,}..., x_n \in \mathbb{R}$, $A_{1,}A_{2,}..., A_n \in .\mathbb{R}$, and

$$y(x; a, b, c) = a + bx + \sum_{i=1}^{n} A_i e^{-\frac{(x-x_i)^2}{2c^2}} \tag{B.1}$$

for $x \in \mathbb{R}$.

Then roughness, $r(a, b, c)$ is

$$r(a, b, c) = \int_{-\infty}^{\infty} (y''(x; a, b, c))^2 dx = \frac{\sqrt{\pi}}{c^3} \sum_{i=1}^{n} \sum_{j=1}^{n} A_i A_j e^{\gamma} \left\{ \frac{3}{4} + 3\gamma + \gamma^2 \right\}, \tag{B.2}$$

where

$$\gamma = \frac{-(x_i - x_j)^2}{4c^2}. \tag{B.3}$$

Proof

Note that

$$\frac{\partial A_i e^{-(\frac{(x-x_i)^2}{2c^2})}}{\partial x} = -\frac{A_i}{c^2} e^{-1/2[\frac{(x-x_i)^2}{2c^2}]} - \frac{A_i}{c^4}(x - x_i)^2 e^{-1/2[\frac{(x-x_i)^2}{2c^2}]},$$

and

$$\frac{\partial^2 A_i e^{-(\frac{(x-x_i)^2}{2c^2})}}{\partial x^2} = \frac{A_i A_j}{c^4} e^{-1/2(\frac{(x-x_i)^2}{c^2})} e^{-1/2(\frac{(x-x_j)^2}{c^2})}$$

$$-\frac{A_i A_j}{c^6} e^{-1/2(\frac{(x-x_i)^2}{c^2})} e^{-1/2(\frac{(x-x_j)^2}{c^2})}(x - x_j)^2$$

$$-\frac{A_i A_j}{c^6} e^{-1/2(\frac{(x-x_i)^2}{c^2})} e^{-1/2(\frac{(x-x_j)^2}{c^2})}(x - x_i)^2$$

$$-\frac{A_i A_j}{c^8} e^{-1/2(\frac{(x-x_i)^2}{c^2})} e^{-1/2(\frac{(x-x_j)^2}{c^2})} (x-x_i)^2 (x-x_j)^2. \tag{B.4}$$

Then roughness for n points is

$$r(a,b,c) = \int_{-\infty}^{\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} (\frac{A_i A_j}{c^4} e^{-1/2(\frac{(x-x_i)^2}{c^2})} e^{-1/2(\frac{(x-x_j)^2}{c^2})}$$

$$-\frac{A_i A_j}{c^6} e^{-1/2(\frac{(x-x_i)^2}{c^2})} e^{-1/2(\frac{(x-x_j)^2}{c^2})} (x-x_j)^2$$

$$-\frac{A_i A_j}{c^6} e^{-1/2(\frac{(x-x_i)^2}{c^2})} e^{-1/2(\frac{(x-x_j)^2}{c^2})} (x-x_i)^2$$

$$\frac{A_i A_j}{c^8} e^{-1/2(\frac{(x-x_i)^2}{c^2})} e^{-1/2(\frac{(x-x_j)^2}{c^2})} (x-x_i)^2 (x-x_j)^2) dx. \tag{B.5}$$

Note that terms that may be separately integrated, and that three general forms appear in Equation (B.5).

First general form

The first general form is $\frac{HJ}{c^4} \int_{-\infty}^{\infty} e^{-1/2(\frac{(x-s)^2}{c^2})} e^{-1/2(\frac{(x-t)^2}{c^2})} dx$, and

$$\frac{HJ}{c^4} \int_{-\infty}^{\infty} e^{-1/2(\frac{(x-s)^2}{c^2})} e^{-1/2(\frac{(x-t)^2}{c^2})} dx = \frac{HJ}{c^4} \int_{-\infty}^{\infty} e^{lx^2+wx+k} dx, \tag{B.6}$$

where $l = \frac{-1}{2c^2}(2)$, $w = \frac{-1}{2c^2}(2s-2t)$, and $k = \frac{-1}{2c^2}(s^2+t^2)$.

Substitute $p = 1/c^2, q = -w/2l = \frac{s+t}{2}$, and $v = k + pq^2$, into Equation (B.6):

$$\int_{-\infty}^{\infty} e^{-1/2(\frac{(x-s)^2}{c^2})} e^{-1/2(\frac{(x-t)^2}{c^2})} dx = \int_{-\infty}^{\infty} e^{lx^2+wx+k} dx = e^v \int_{-\infty}^{\infty} e^{-p(x-q)^2} dx. \qquad \text{(B.7)}$$

Note applying the definition of a Gaussian probability density, $\frac{1}{\sqrt{2\pi d}} \int_{-\infty}^{\infty} e^{-\frac{(x-s)^2}{2d^2}} dx = 1$, yields the following progression:

$$e^r \int_{-\infty}^{\infty} e^{-p(x-q)^2} dx = e^v \sqrt{\pi} c. \qquad \text{(B.8)}$$

Thus

$$\frac{HJ}{c^4} \int_{-\infty}^{\infty} e^{-1/2(\frac{(x-s)^2}{c^2})} e^{-1/2(\frac{(x-t)^2}{c^2})} dx = \frac{HJ}{c^4} e^v \sqrt{\pi}, \qquad \text{(B.9)}$$

where $v = k + pq^2$ and $k = \frac{-1}{2c^2}(s^2 + t^2)$, $p = 1/c^2$, and $q = \frac{s+t}{2}$

Second general form

$$\frac{HJ}{c^4} \int_{-\infty}^{\infty} e^{-1/2(\frac{(x-s)^2}{c^2})} e^{-1/2(\frac{(x-t)^2}{c^2})} (x-t)^2 dx \qquad \text{(B.10)}$$

Similar to Equation (B.6), let

$$\int_{-\infty}^{\infty} e^{-1/2(\frac{(x-s)^2}{c^2})} e^{-1/2(\frac{(x-t)^2}{c^2})} (x-t)^2 dx = \int_{-\infty}^{\infty} e^{lx^2+wx+k} (x-t)^2 dx. \qquad \text{(B.11)}$$

Then, similar to Equation (B.7), $\int_{-\infty}^{\infty} e^{lx^2+wx+k}(x-t)^2 dx = e^r \int_{-\infty}^{\infty} e^{-p(x-q)^2}(x-t)^2 dx$.
Note that $e^r \int_{-\infty}^{\infty} e^{-p(x-q)^2}(x-t)^2 dx =$
$t^2 e^r \int_{-\infty}^{\infty} e^{-p(x-q)^2} dx - 2te^v \int_{-\infty}^{\infty} x e^{-p(x-q)^2} dx + e^r \int_{-\infty}^{\infty} x^2 e^{-p(x-q)^2} dx$.

Similar to Equation (B.8), $t^2 e^v \int_{-\infty}^{\infty} e^{-p(x-q)^2} dx = t^2 e^v \sqrt{\pi} c$.

Substitute $z = x - q$. Then $x = z + q$, and
$\int_{-\infty}^{\infty} x e^{-p(x-q)^2} dx = \int_{-\infty}^{\infty} (z+q) e^{-p(z)^2} dx = \int_{-\infty}^{\infty} z e^{-p(z)^2} dz + q \int_{-\infty}^{\infty} e^{-p(z)^2} dz$.

Note that $\int_{-\infty}^{\infty} ze^{-p(z)^2}dz = 0$.

(Recall that $p = 1/c^2$.)

Then $-2te^v \int_{-\infty}^{\infty} xe^{-p(x-q)^2}dx = -2te^v(0 + q\sqrt{2\pi}\frac{c}{\sqrt{2}}) = -2te^v q\sqrt{\pi}c$.

Similar to above, let $z = x - q$, and $\int_{-\infty}^{\infty} x^2 e^{-p(x-q)^2}dx = \int_{-\infty}^{\infty}(z+q)^2 e^{-p(z)^2}dz$.

Then

$\int_{-\infty}^{\infty}(z+q)^2 e^{-p(z)^2}dz = \int_{-\infty}^{\infty}(z^2 + 2qz + q^2)e^{-p(z)^2}dz =$
$\int_{-\infty}^{\infty} z^2 e^{-p(z)^2}dz + \int_{-\infty}^{\infty} 2qze^{-p(z)^2}dz + \int_{-\infty}^{\infty} q^2 e^{-p(z)^2}dz$.

Note that $\int_{-\infty}^{\infty} 2qze^{-p(z)^2}dz = 0$, and similar to Equation (B.8),
$\int_{-\infty}^{\infty} q^2 e^{-p(z)^2}dz = q^2\sqrt{\pi}c$.

$\int_{-\infty}^{\infty} z^2 e^{-p(z)^2}dz = (\frac{c}{2})^2\sqrt{2\pi(\frac{c}{2})^2} = \frac{c^3}{2}\sqrt{\pi}$.

Thus
$\frac{HJ}{c^4}\int_{-\infty}^{\infty} e^{-1/2(\frac{(x-s)^2}{c^2})}e^{-1/2(\frac{(x-t)^2}{c^2})}(x-t)^2dx = t^2e^v\sqrt{\pi}c - 2te^v q\sqrt{\pi}c + e^v\frac{c^3}{2}\sqrt{\pi} + e^v q^2\sqrt{\pi}c$.

Note that factoring:

$t^2e^v\sqrt{\pi}c - 2te^v q\sqrt{\pi}c + e^v\frac{c^3}{2}\sqrt{\pi} + e^v q^2\sqrt{\pi}c = e^v\sqrt{\pi}c[t^2 - 2tq + \frac{c^2}{2} + q^2]$.

Thus,

$$\frac{HJ}{c^4}\int_{-\infty}^{\infty} e^{-1/2(\frac{(x-s)^2}{c^2})}e^{-1/2(\frac{(x-t)^2}{c^2})}(x-t)^2dx = e^v\sqrt{\pi}c[t^2 - 2tq + \frac{c^2}{2} + q^2]. \quad \text{(B.12)}$$

Third General Form

A similar progression yields

$$\frac{HJ}{c^4}\int_{-\infty}^{\infty} e^{-1/2(\frac{(x-s)^2}{c^2})}e^{-1/2(\frac{(x-t)^2}{c^2})}(x-x)^2(x-t)^2dx =$$

$$ce^v\sqrt{\pi}\{q^4+c^2(3q^2)+\frac{3c^4}{4}+[(-2s-2t)(q^3+\frac{3qc^2}{2})]$$

$$+[(s^2+t^2+4st)(q^2+(\frac{c^2}{2}))].$$

$$+[(q)(-2st^2-2s^2t)]+[s^2t^2]\}. \tag{B.13}$$

Applying Equations (B.9), (B.12), and (B.13) to roughness formula:

$$r(a,b,c)=\{\frac{\sqrt{\pi}}{c^3}\sum_{i=1}^{n}\sum_{j=1}^{n}A_iA_je^{\frac{v-q^2}{c^2}}[1-\frac{2s^2}{c^2}+\frac{4sq}{c^2}-\frac{2q^2}{c^2}-1$$

$$+\frac{q^4}{c^4}+\frac{3q^2}{c^2}+\frac{3}{4}+\frac{-2sq^3}{c^4}+\frac{-3qs}{c^2}+\frac{-2tq^3}{c^4}+\frac{-3qt}{c^2}+\frac{s^2q^2}{c^4}+\frac{s^2}{2c^2}+\frac{t^2q^2}{c^4}+\frac{t^2}{2c^2}$$

$$+\frac{4stq^2}{c^4}+\frac{2st}{c^2}+\frac{-2st^2q}{c^4}+\frac{-2s^2tq}{c^4}+\frac{s^2t^2}{c^4}]\}^{1/2}. \tag{B.14}$$

Note that the following terms

$$\frac{q^4}{c^4}+\frac{-2sq^3}{c^4}+\frac{-2tq^3}{c^4}+\frac{s^2q^2}{c^4}+\frac{t^2q^2}{c^4}+\frac{4stq^2}{c^4}+\frac{-2st^2q}{c^4}+\frac{-2s^2tq}{c^4}+\frac{s^2t^2}{c^4}, \tag{B.15}$$

factor as follows.

The second and third terms of Equation (B.15) are

$$\frac{-2sq^3}{c^4} + \frac{-2tq^3}{c^4} = \frac{-2q^3(s+t)}{c^4} = \frac{-4(s+t)q^3}{2c^4} = \frac{-4q^4}{c^4}, \tag{B.16}$$

and thus the first second and third terms of Equation (B.15) are

$$\frac{q^4}{c^4} + \frac{-2sq^3}{c^4} + \frac{-2tq^3}{c^4} = \frac{-3q^4}{c^4}. \tag{B.17}$$

Note that the seventh and eighth terms of Equation (B.15) are

$$\frac{-2st(s+t)q}{c^4} = \frac{-4st(s+t)q}{2c^4} = \frac{-4stq^2}{c^4}, \tag{B.18}$$

and thus the sixth, seventh and eighth terms in Equation (B.15) are:

$$\frac{4stq^2}{c^4} + \frac{-4stq^2}{c^4} = 0. \tag{B.19}$$

Note that Equation (B.15) is equal to

$$c^4 \text{ terms} = \frac{-3q^4}{c^4} + \frac{s^2q^2}{c^4} + \frac{t^2q^2}{c^4} + \frac{s^2t^2}{c^4} \tag{B.20}$$

Note that the second and third terms in Equation (B.20) simplify to:

$$\frac{s^2q^2}{c^4} + \frac{t^2q^2}{c^4} = \frac{q^2(s^2+t^2)}{c^4} = \frac{q^2[(s+t)^2 - 2st]}{c^4} = \frac{4q^2[(s+t)^2 - 2st]}{4c^4}, \tag{B.21}$$

and that

$$\frac{4q^2[(s+t)^2 - 2st]}{4c^4} = \frac{4q^2(s+t)^2}{4c^4} + \frac{4q^2(-2st)}{4c^4}. \tag{B.22}$$

Note that

$$\frac{4q^2(s+t)^2}{4c^4} + \frac{4q^2(-2st)}{4c^4} = \frac{4q^4}{c^4} + \frac{4q^2(-2st)}{4c^4} \tag{B.23}$$

Therefore, Equation (B.20) simplifies to:

$$c^4 \text{ terms} = \frac{-3q^4}{c^4} + \frac{4q^4}{c^4} + \frac{4q^2(-2st)}{4c^4} + \frac{s^2t^2}{c^4} \tag{B.24}$$

From above equation:

$$c^4 \text{ terms} = \frac{q^4}{c^4} + \frac{-2q^2st}{c^4} + \frac{s^2t^2}{c^4} \tag{B.25}$$

Substitute $v = st$.

$$c^4 \text{ terms} = \frac{q^4}{c^4} + \frac{-2q^2v}{c^4} + \frac{v^2}{c^4} = \frac{(q^2 - v)^2}{c^4} \tag{B.26}$$

Next, examine the 11 terms with denominator of c^2 :

$$c^2 \text{ terms} = \frac{-t^2}{c^2} + \frac{2tq}{c^2} + \frac{-2q^2}{c^2} + \frac{-s^2}{c^2} + \frac{2sq}{c^2} + \frac{3q^2}{c^2} + \frac{-3qs}{c^2} + \frac{-3qt}{c^2} + \frac{s^2}{2c^2} + \frac{t^2}{2c^2} + \frac{2st}{c^2} \tag{B.27}$$

Note that the seventh and eighth terms in Equation (B.27) simplify to

$$\frac{-3qs}{c^2} + \frac{-3qt}{c^2} = \frac{-3q(s+t)}{c^2} = \frac{-6q(s+t)}{2c^2} = \frac{-6q^2}{c^2} \tag{B.28}$$

Combining the third, sixth, seventh, and eighth terms in Equation (B.27):

$$\frac{-2q^2}{c^2} + \frac{3q^2}{c^2} + \frac{-6q^2}{c^2} = \frac{-5q^2}{c^2} \tag{B.29}$$

Now, Equation (B.27) is simplified as

$$c^2 \text{ terms} = \frac{-t^2}{c^2} + \frac{2tq}{c^2} + \frac{-5q^2}{c^2} + \frac{-s^2}{c^2} + \frac{2sq}{c^2} + \frac{s^2}{2c^2} + \frac{t^2}{2c^2} + \frac{2st}{c^2} \tag{B.30}$$

Note that the second and fifth terms in Equation (B.30) are:

$$\frac{2tq}{c^2} + \frac{2sq}{c^2} = \frac{2q(s+t)}{c^2} = \frac{4q(s+t)}{2c^2} = \frac{4q^2}{c^2}, \tag{B.31}$$

and thus Equation (B.30) simplifies to:

$$c^2 \text{ terms} = \frac{-t^2}{c^2} + \frac{-q^2}{c^2} + \frac{-s^2}{c^2} + \frac{s^2}{2c^2} + \frac{t^2}{2c^2} + \frac{2st}{c^2} \tag{B.32}$$

Combining the first, third, fourth, and fifth terms of Equation (B.32):

$$\frac{-t^2}{c^2} + \frac{-s^2}{c^2} + \frac{s^2}{2c^2} + \frac{t^2}{2c^2} = \frac{-t^2}{2c^2} + \frac{-s^2}{2c^2} \tag{B.33}$$

Now, Equation (B.32) simplifies to:

$$c^2 \text{ terms} = \frac{-t^2 - s^2}{2c^2} - \frac{q^2}{c^2} + \frac{2st}{c^2} \tag{B.34}$$

Replace $\frac{(s+t)}{2}$ for q in the above equation, and use a common denominator to obtain:

$$c^2 \text{ terms} = \frac{-2t^2 - 2s^2}{4c^2} - \frac{(s+t)^2}{4c^2} + \frac{8st}{4c^2} \tag{B.35}$$

Therefore,

$$c2 \text{ terms} = \frac{-2t^2 - 2s^2 - (s^2 + 2st + t^2) + 8st}{4c^2} = \frac{-2t^2 - 2s^2 - s^2 - 2st - t^2 + 8st}{4c^2}, \tag{B.36}$$

$$\frac{-2t^2 - 2s^2 - s^2 - 2st - t^2 + 8st}{4c^2} = \frac{-3t^2 - 3s^2 + 6st}{4c^2}, \tag{B.37}$$

and

$$c2 \text{ terms} = \frac{-3(t^2 + s^2 - 2st)}{4c^2} = \frac{-3((s+t)^2 - 4st)}{4c^2}$$

$$= \frac{-3q^2}{c^2} + \frac{12v}{4c^2} = \frac{-3q^2}{c^2} + \frac{3v}{c^2} = \frac{3(v - q^2)}{c^2}. \tag{B.38}$$

Combining the c^4, c^2, and constant terms, and inserting into the roughness formula:

$$r(a,b,c) = \left\{ \frac{\sqrt{\pi}}{c^3} \sum_{i=1}^{n} \sum_{j=1}^{n} A_i A_j e^{\frac{v-q^2}{c^2}} \left\{ \frac{3}{4} + \frac{(3)(v-q^2)}{c^2} + \frac{(v-q^2)^2}{c^4} \right\} \right\}^{1/2} \tag{B.39}$$

Substitute $\gamma = \frac{v-q^2}{c^2}$.

Note that

$$\gamma = \frac{v-q^2}{c^2} = \frac{v}{c^2} - \frac{q^2}{c^2} = \frac{st}{c^2} - \frac{(s+t)^2}{4c^2} = \frac{4st-(s+t)^2}{4c^2} = \frac{4st-(s^2+2st+t^2)}{4c^2}, \tag{B.40}$$

and

$$\gamma = \frac{4st-s^2-2st-t^2}{4c^2} = \frac{-(s^2-2st+t^2)}{4c^2} = \frac{-(s-t)^2}{4c^2} \tag{B.41}$$

Thus,

$$r(a,b,c) = \int (y''(x;a,b,c))^2 dx = \frac{\sqrt{\pi}}{c^3} \sum_{i=1}^{n} \sum_{j=1}^{n} A_i A_j e^{\gamma} \left\{ \frac{3}{4} + 3\gamma + \gamma^2 \right\}, \tag{B.42}$$

where

$$\gamma = \frac{-(x_i - x_j)^2}{4c^2}. \tag{B.43}$$

Appendix C. ROC Curve and CEG Curve Probability Density and Confidence Interval Software

Appendix C-1 details code [Parker, 2005] that computes median estimates of ROC curves and AUC values, with confidence intervals, for any set of target and non-target input score samples, assuming beta target and non-target densities. Appendix C-2 is identical in purpose, except that it assumes two-beta mixture target and non-target densities. These appendices contain instructions for additional code that assumes target and non-target densities with fixed, user-specified parameters. This additional code generates many sets of representative target and non-target samples from the fixed densities, and it provides corresponding ROC curve coverage accuracies. Appendix C-3 and C-4 describe code identical in purpose to C-1 and C-2, but for CEG curves and RSD values. The end of Section 5 compares the beta and two-beta density approaches; the principal approach applied in the research reported here is the single beta model. The code for each of the Matlab files that comprise the user interface is also provided here. The remaining Matlab files are functions that are called upon execution of the user interface.

Appendix C-1
ROC curve /AUC value Estimation and Confidence Interval
Matlab Instructions
Beta Density Target and Non-target Model

A. Provide a set of target samples, non-target samples, and confidence bound value. Then compute the confidence intervals based on these samples.

1. Place the following files into a common directory.
   (For example: c:\matlab_sv12\work\)

   beta_mean_w_a_b_r.m
   conditioned_calc_2_r.m
   find_max_variance_r.m
   get_aurc_val_r.m
   get_density_vals_r.m
   get_grid_points_closest_r.m
   get_grid_points_n_closest_r.m
   get_grid_points_r.m
   get_pd_pfa_matrix_10_r.m
   get_pd_pfa_pairs_pdfs2_r.m
   high_low_grid_weight_r.m
   mean_variance_to_pdf_2_r.m
   pd_pfa_from_mean_std_r.m
   pfa_pd_to_hundredths_r.m
   script_for_samples_r.m
   uni_pdf_for_samples_r

2. Add the common directory to the Matlab path by using 'File / Set Path' option in Matlab menu, if the directory is not already in the path.

3. Execute the following in Matlab. An example is contained in 'script_for_samples_r.m'.

Enter (or load) a vector of target scores into the variable 'new_target_scores'.
Enter (or load) a vector of nontarget scores into the variable 'new_nontarget_scores'.

Execute the following matlab code to produce ROC curve and AUC value estimates with confidence intervals:

uni_pdf_for_samples_r(new_target_scores,new_nontarget_scores,.95);

Replace .95 by alternate confidence interval coverage if desired (e.g. .90, .80).

To obtain the upper and lower confidence interval limit values for false alarm probabilities 0, .01, …, .99, 1, (rather than an on-screen plot), execute the following:

[ci_median, ci_upper, ci_lower, auc_median, auc_upper, auc_lower] = …
uni_pdf_for_samples_r(new_target_scores,new_nontarget_scores,bound_value);

ci_median – ROC curve estimate
ci_upper - Upper ROC curve confidence interval contour
ci_lower - Lower ROC curve confidence interval contour
auc_median - AUC value estimate
auc_upper - Upper AUC value confidence interval estimate
auc_lower - Lower AUC value confidence interval estimate

B. Generate many sets of samples for selected underlying target and nontarget densities, and then obtain confidence intervals and estimates for the ROC / AUC for each set of samples and compute confidence interval accuracy (e.g. alpha) among all sets. This process assumes a single beta model for target and non-target.

1. Place the following files into a common directory.

   For example: c:\matlab_sv12\work\roc\

   beta_mean_w_a_b_r.m
   conditioned_calc_2_r.m
   find_max_variance_r.m
   generic_rnd.m
   get_aurc_val_r.m
   get_density_vals_r.m
   get_grid_points_closest_r.m
   get_grid_points_n_closest_r.m
   get_grid_points_r.m
   get_pd_pfa_matrix_10_r.m
   get_pd_pfa_pairs_pdfs2_r.m
   high_low_grid_weight_r.m
   mean_variance_to_pdf_2_r.m
   pd_pfa_from_mean_std_r.m
   pfa_pd_to_hundredths_r.m
   run_choose_sample.m
   run_Uu_aurc_95_r.m
   sample_gen_uni_test_r.m
   sample_gen_user_input_r.m
   script_ROC_AUC_CIs_with_coverage_accuracy.m
   uni_pdf_aurc_95_r.m

2. Add the common directory to the Matlab path by using 'File / Set Path' option in Matlab menu.

3. Open the File 'script_ROC_AUC_CIs_with_coverage_accuracy.m'.

   Lines 7-18. Specify number of target samples, number of non-target samples, specify a beta density by mean and variance of assumed target beta density, mean and variance of assumed non-target beta density number of runs, or provide any density form as input (Lines 23-24 provide an example).

   Evaluate Lines 1 through 88. [Note in Matlab this can be achieved by highlighting these line. Then right click to obtain a menu. Then choose 'Evaluate Selection'.] After each run, the full set of results are saved in Line 88.

   ROC curve for a single run with confidence intervals:

   a. Form a plot of estimated ROC with confidence intervals [with true ROC] by Evaluating Lines 92-105. Note that run_number on line 92 may be adjusted to any run among the set specified in step 3 above.

   Obtain coverage for the full set of runs by Evaluating Lines 111-176. The mean alpha for AUC over many runs is displayed at the top of the plot.

```
 1
 2 % Example target and non-target scores
 3
 4 new_target_scores =  [0.6876; 0.7125; 0.7400; 0.7491; 0.7496; 0.7595; 0.7601; 0.7643; 0.7710; 0.7724; 0.7732; 0.7904;↙
0.8010; 0.8100;...
 5         0.8171; 0.8200; 0.8218; 0.8293; 0.8300; 0.8431; 0.8464; 0.8464; 0.8534; 0.8600; 0.8756; 0.8831; 0.8941;↙
0.8991; 0.9055; 0.9256];
 6
 7 new_nontarget_scores = [0.6204; 0.6351; 0.6584; 0.6662; 0.6750; 0.6782; 0.6861; 0.6902; 0.7008; 0.7009; 0.7013;↙
0.7037; 0.7054; 0.7181;...
 8         0.7204; 0.7218; 0.7221; 0.7222; 0.7261; 0.7274; 0.7296; 0.7362; 0.7444; 0.7447; 0.7556; 0.7585; 0.7602;↙
0.7808; 0.7856; 0.7956];
 9
10
11
12 % load new_target_scores;
13 % load new_nontarget_scores;
14
15
16
17 % Specify the desired coverage:
18 bound_value = .95;
19
20
21
22 % To generate a ROC curve:
23 uni_pdf_for_samples_r(new_target_scores,new_nontarget_scores,bound_value);
24
25
26
27 % To obtain the values for the ROC curve estimate, confidence interval, and AUC estimate and confidence values:
28 [ci_median, ci_upper, ci_lower, auc_median, auc_upper, auc_lower] = uni_pdf_for_samples_r(new_target_scores,↙
new_nontarget_scores,bound_value);
29
30
31
32 % To plot the ROC curve estimate with confidence intervals:
33 plot(0:.01:1,ci_median,'k-.');
```

```
34 hold on;
35 plot(0:.01:1,ci_upper,'k:');
36 hold on;
37 plot(0:.01:1,ci_lower,'k:');
38 axis equal;
39 axis([-.1 1.1 -.1 1.1]);
40 xlabel('false alarm probability');
41 ylabel('correct detection probability');
42 title(['auc est. = ',num2str(auc_median),' auc lower c.i. = ',num2str(auc_lower),' auc upper c.i. = ',num2str↙
(auc_upper)]);
43
44
45
46
47
```

```

% script_ROC_AUC_CIs_with_coverage_accuracy.m

clear all;
target_density = zeros(1000,1);
nontarget_density = zeros(1000,1);


number_of_target_samples = 200;
number_of_nontarget_samples = 200;

% Enter the beta density parameters
mean_target_gen_1 = .599;
variance_target_gen_1 = .021;
mean_nontarget_gen_1 = .479;
variance_nontarget_gen_1 = .023;

bound_val = .95;

number_of_runs = 100;

% optional:
% Enter target density and nontarget densities in place of above beta parameter values
% % A beta pdf is used here as an example.
% target_density = betapdf(0:.001:1,1,3);
% nontarget_density = betapdf(0:.001:1,3,1);


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


length_run_number = 1;
plot_1_uni_ci_median_vector = zeros(length_run_number,101);
plot_1_uni_ci_upper_vector = zeros(length_run_number,101);
plot_1_uni_ci_lower_vector = zeros(length_run_number,101);
pfa_1_vector_true_curve_vector = zeros(length_run_number,1001);
```

```
 39 pd_1_vector_true_curve_vector = zeros(length_run_number,1001);
 40 alpha_1_diff_many_runs = zeros(length_run_number,1);
 41
 42 aurc_1_median_many_runs = zeros(length_run_number,1);
 43 aurc_1_upper_many_runs = zeros(length_run_number,1);
 44 aurc_1_lower_many_runs = zeros(length_run_number,1);
 45 alpha_1_aurc_many_runs = zeros(length_run_number,1);
 46 true_1_aurc_many_runs = zeros(length_run_number,1);
 47
 48 uni_1_conf_check_val_many_runs = zeros(1,length_run_number);
 49 mse_1_uni_many_runs = zeros(1,length_run_number);
 50 test_1_uni_many_runs = zeros(1,length_run_number);
 51
 52 run_number_start = 1;
 53 run_number_end = number_of_runs;
 54
 55 for run_number = run_number_start:run_number_end
 56
 57
 58     if max(target_density) == 0
 59     [plot_uni_ci_median,plot_uni_ci_upper,plot_uni_ci_lower,uni_conf_check_val,pfa_vector_true_curve,↙
pd_vector_true_curve,mse_uni,test_uni,alpha_diff,aurc_median,aurc_upper,aurc_lower,alpha_aurc,true_aurc] =↙
run_Uu_aurc_95_r(run_number,number_of_target_samples,number_of_nontarget_samples,...
 60         mean_target_gen_1,variance_target_gen_1,mean_nontarget_gen_1,variance_nontarget_gen_1,bound_val);
 61     end;
 62
 63     if max(target_density) > 0
 64     [plot_uni_ci_median,plot_uni_ci_upper,plot_uni_ci_lower,uni_conf_check_val,pfa_vector_true_curve,↙
pd_vector_true_curve,mse_uni,test_uni,alpha_diff,aurc_median,aurc_upper,aurc_lower,alpha_aurc,true_aurc] =↙
run_choose_sample(run_number,number_of_target_samples,number_of_nontarget_samples,...
 65         target_density,nontarget_density,bound_val);
 66     end;
 67
 68     aurc_1_median_many_runs(run_number) = aurc_median;
 69     aurc_1_upper_many_runs(run_number) = aurc_upper;
 70     aurc_1_lower_many_runs(run_number) = aurc_lower;
 71     alpha_1_aurc_many_runs(run_number) = alpha_aurc;
 72     true_1_aurc_many_runs(run_number) = true_aurc;
```

```
 73
 74     uni_1_conf_check_val_many_runs(run_number) = uni_conf_check_val;
 75
 76     mse_1_uni_many_runs(run_number) = mse_uni;
 77
 78     test_1_uni_many_runs(run_number) = test_uni;
 79     alpha_1_diff_many_runs(run_number) = alpha_diff;
 80
 81     plot_1_uni_ci_median_vector(run_number,:) = plot_uni_ci_median;
 82     plot_1_uni_ci_upper_vector(run_number,:) = plot_uni_ci_upper;
 83     plot_1_uni_ci_lower_vector(run_number,:) = plot_uni_ci_lower;
 84     pfa_1_vector_true_curve_vector(run_number,:) = pfa_vector_true_curve;
 85     pd_1_vector_true_curve_vector(run_number,:) = pd_vector_true_curve;
 86
 87     save store_confidence_intervals_for_each_run
 88
 89 end;
 90
 91
 92 % Plot the ROC curve for a selected run.
 93 run_number = 1; % This may be changed to particular runs.
 94
 95 plot(0:.01:1,plot_1_uni_ci_median_vector(run_number,:),'k-.');
 96 hold on;
 97 plot(0:.01:1,plot_1_uni_ci_upper_vector(run_number,:),'k:');
 98 hold on;
 99 plot(0:.01:1,plot_1_uni_ci_lower_vector(run_number,:),'k:');
100 hold on;
101 plot(pfa_1_vector_true_curve_vector(run_number,:),pd_1_vector_true_curve_vector(run_number,:),'k');
102 legend('median','upper ci','lower ci','truth (underlying)');
103 title(['alpha est. = ',num2str(1-uni_1_conf_check_val_many_runs(run_number)),' run # = ',num2str(run_number),' #↙
target samples = ',num2str(number_of_target_samples),' # of nontarget samples = ',num2str↙
(number_of_nontarget_samples)]);
104 axis equal;
105 axis([-.1 1.1 -.1 1.1]);
106 title(' ');
107
108
```

```
% alpha estimate per point

[pfa_true_curve_mod,pd_true_curve_mod] = pfa_pd_to_hundredths_r(pfa_1_vector_true_curve_vector(1,:),↙
pd_1_vector_true_curve_vector(1,:))

vector_check = [run_number_start:1:run_number_end]

tot_runs_to_use = length(vector_check);

check_val_vector = zeros(101,1);
length_val_vector = zeros(101,1);

for index_a = 1:101
    for index_b_wild = 1:tot_runs_to_use
        index_b = vector_check(index_b_wild);
        truth_compare_now = pd_true_curve_mod(index_a);
        % metz process
        low_compare_now = plot_1_uni_ci_lower_vector(index_b,index_a)
        up_compare_now = plot_1_uni_ci_upper_vector(index_b,index_a)
        length_now = up_compare_now-low_compare_now;
        check_val = 1;
        if low_compare_now <= truth_compare_now
            if up_compare_now >= truth_compare_now
                check_val = 0;
            end;
        end;
        check_val_vector(index_a) = check_val_vector(index_a) + check_val;
        length_val_vector(index_a) = length_val_vector(index_a) + length_now;
    end;
    check_val_vector(index_a) = check_val_vector(index_a)./tot_runs_to_use;
    length_val_vector(index_a) = length_val_vector(index_a)./tot_runs_to_use;
end;



test_pdf_vector = zeros(101,tot_runs_to_use);
cum_test_pdf_vector = zeros(size(test_pdf_vector));
Y_min = zeros(100,1);
```

```
146 Y_max = zeros(100,1);
147 min_error_vector = zeros(100,1);
148 max_error_vector = zeros(100,1);
149
150 for index = 2:100
151     test_pdf = binopdf(1:1:length(vector_check),length(vector_check),check_val_vector(index));
152     test_pdf_vector(index,:) = test_pdf;
153     for index_2 = 1:tot_runs_to_use
154         cum_test_pdf_vector(index,index_2) = sum(test_pdf_vector(index,1:index_2));
155     end;
156     [Y_min(index),I_min(index)] = min(abs(.05-cum_test_pdf_vector(index,:)));
157     [Y_max(index),I_max(index)] = min(abs(.95-cum_test_pdf_vector(index,:)));
158     min_error_vector(index) = I_min(index)./tot_runs_to_use;
159     max_error_vector(index) = I_max(index)./tot_runs_to_use;
160     index = index
161 end;
162
163 x_vals = 0:.01:1;
164 for index = 2:100
165     errorbar(x_vals(index),100.*(1-check_val_vector(index)),100.*(1-check_val_vector(index))-100.*(1-↙
min_error_vector(index)),100.*(1-max_error_vector(index))-100.*(1-check_val_vector(index)),'k');
166     hold on;
167 end;
168
169 line([-2 2],[95 95],'Color','k');
170 hold on;
171 plot(.01:.01:.99,(100.*(1-check_val_vector(2:100))));
172 hold on;
173 mean_alpha = mean(1-uni_conf_check_val_many_runs);
174 %pbaspect([1 3 2]);
175 title(['mean auc alpha = ',num2str(mean_alpha)]);
176 axis([-.05 1.05 0 102]);
```

Appendix C-2
ROC curve /AUC value Estimation and Confidence Interval Matlab Instructions
Two-Beta Mixture Target and Non-Target Density Model

A. Provide a set of target samples, non-target samples, and confidence bound value. Then compute the confidence intervals based on these samples (assumes a two-beta mixture model).

1. Place the following files into a common directory.
   (For example: c:\matlab_sv12\work\)

   beta_mean_w_ab_2br.m
   combine_beta_pdf_2br.m
   conditioned_calc_2_2br.m
   find_max_variance_2br.m
   get_aurc_val_2br.m
   get_pd_pfa_matrix_10_2br.m
   get_pd_pfa_pairs_pdfs2_2br.m
   mixture_pdf_2br.m
   pfa_pd_to_hundredths_2br.m
   rand_two_beta_density_2br.m
   roc_from_density_2br.m
   sample_gen_bimodal_2br.m
   two_beta_script_for_given_samples_2br.m
   two_beta_roc_truth_not_known_2br.m

2. Add the common directory to the Matlab path by using 'File / Set Path' option in Matlab menu, if the directory is not already in the path.

3. Execute the following in Matlab. An example is contained in 'two_beta_script_for_given_samples_2br.m'.

Enter (or load) a vector of target scores into the variable 'new_target_scores'.
Enter (or load) a vector of nontarget scores into the variable 'new_nontarget_scores'.

Execute the following matlab code to produce ROC curve and AUC value estimates with confidence intervals:

two_beta_roc_truth_not_known(new_target_scores,new_nontarget_scores,10000,.95);

Replace 10000 by the desired number of random draws (lower numbers of draws decrease computational time). An approach is to begin with a low number of draws and gradually increase until convergence of confidence interval solution is observed.
Replace .95 by alternate confidence interval coverage if desired (e.g. .90, .80).

To obtain the upper and lower confidence interval limit values for false alarm probabilities 0, .01, …, .99, 1, (rather than an on-screen plot), execute the following:
[ci_median, ci_upper, ci_lower, auc_median, auc_upper, auc_lower] = …
two_beta_roc_truth_not_known(new_target_scores,new_nontarget_scores,10000,.95);

ci_median –	ROC curve estimate
ci_upper -	Upper ROC curve confidence interval contour
ci_lower -	Lower ROC curve confidence interval contour
auc_median -	AUC value estimate
auc_upper -	Upper AUC value confidence interval estimate
auc_lower -	Lower AUC value confidence interval estimate

B. Generate many sets of samples for selected underlying target and nontarget densities, and then obtain confidence intervals and estimates for the ROC curve / AUC value for each set of samples and compute confidence interval accuracy (e.g. alpha) among all sets. This process assumes a two-beta mixture model for target and non-target.

1. Place the following files into a common directory.

   For example: c:\matlab_sv12\work\roc\

   beta_mean_w_ab_2br.m
   combine_beta_pdf_2br.m
   conditioned_calc_2_2br.m
   find_max_variance_2br.m
   get_aurc_val_2br.m
   get_pd_pfa_matrix_10_2br.m
   get_pd_pfa_pairs_pdfs2_2br.m
   mixture_pdf_2br.m
   pfa_pd_to_hundredths_2br.m
   rand_two_beta_density_2br.m
   roc_from_density_2br.m
   sample_gen_bimodal_2br.m
   two_beta_script_for_many_runs_2br.m
   twobeta_run_nonempirical_2br.m
   two_beta_unipdf_aurc_1000_2br.m

2. Add the common directory to the Matlab path by using 'File / Set Path' option in Matlab menu.

3. Open the File 'two_beta_script_for_many_runs_2br.m'.

   Lines 4-22. Specify number of target samples, number of non-target samples, specify the five parameters for the target density for a two-beta mixture model (two means, two standard deviations, and a ratio), the five parameters for the non-target density, the number of random_draws desired (e.g. 2000), confidence interval desired (ci_range; example is .90 for 90% confidence intervals), and the number of test runs (number_of_runs; example is 100 if 100 test runs are desired). An example is provided; change these values as desired.

   Evaluate Lines 1 through 77. [Note in Matlab this can be achieved by highlighting these lines. Then right click to obtain a menu. Then choose 'Evaluate Selection'.] After each run, the full set of results are saved in Line 88.

   ROC curve for a single run with confidence intervals:

   a. Form a plot of estimated ROC with confidence intervals [with true ROC] by Evaluating Lines 78-93. Note that run_number on line 79 may be adjusted to any run among the set specified in step 3 above.

   b. Obtain coverage for the full set of runs by Evaluating Lines 100-179. The mean alpha for AUC over many runs is displayed at the top of the plot.

```
 1
 2 new_target_scores = [0.7580; 0.7588; 0.7618; 0.7650; 0.7674; 0.7794; 0.7870; 0.7935;...
 3 0.7959; 0.8027; 0.8035; 0.8043; 0.8049; 0.8066; 0.8141; 0.8183; 0.8189; 0.8202;...
 4 0.8262; 0.8328; 0.8479; 0.8487; 0.8505; 0.8637; 0.8701; 0.8718; 0.8753; 0.8831;...
 5 0.8880; 0.8953];
 6
 7 new_nontarget_scores = [0.5988; 0.6331; 0.6396; 0.6642; 0.6652; 0.6688; 0.6734;...
 8 0.6736; 0.6764; 0.6802; 0.6954; 0.6990; 0.7024; 0.7059; 0.7109; 0.7126; 0.7132;...
 9 0.7162; 0.7185; 0.7196; 0.7231; 0.7286; 0.7347; 0.7362; 0.7644; 0.7647; 0.7785;...
10 0.7798; 0.7947; 0.8030];
11
12 ci_range = .90;
13 number_of_random_draws = 2000;
14
15 [uni_value_1, uni_value_2, uni_value_3, uni_value_4, uni_value_5, uni_value_6] = twobeta_roc_truth_not_known_2br↙
(new_target_scores,new_nontarget_scores,number_of_random_draws,ci_range);
16
17 uni_ci_median = uni_value_1;
18 uni_ci_upper = uni_value_2;
19 uni_ci_lower = uni_value_3;
20 aurc_median = uni_value_4;
21 aurc_upper = uni_value_5;
22 aurc_lower = uni_value_6;
23
24 plot(0:.01:1,uni_ci_median,'k-.');
25 hold on;
26 plot(0:.01:1,uni_ci_upper,'k:');
27 hold on;
28 plot(0:.01:1,uni_ci_lower,'k:');
29 hold on;
30 axis equal;
31 axis([-.1 1.1 -.1 1.1]);
32
33
34
```

```
 1
 2 clear all;
 3
 4 number_of_target_samples = 300;
 5 number_of_nontarget_samples = 250;
 6
 7 % First obtain the underlying target and clutter density:
 8 mean_target_1 = .8;
 9 std_target_1 = .08;
10 mean_target_2 = .6;
11 std_target_2 = .1;
12 ratio_target = .5;
13
14 mean_nontarget_1 = .6;
15 std_nontarget_1 = .08;
16 mean_nontarget_2 = .4;
17 std_nontarget_2 = .1;
18 ratio_nontarget = .3;
19
20 number_of_random_draws = 2000;
21 ci_range = .90;
22 number_of_runs = 1;
23
24 length_run_number = 1;
25 plot_uni_ci_median_vector = zeros(length_run_number,101);
26 plot_uni_ci_upper_vector = zeros(length_run_number,101);
27 plot_uni_ci_lower_vector = zeros(length_run_number,101);
28 pfa_vector_true_curve_vector = zeros(length_run_number,101);
29 pd_vector_true_curve_vector = zeros(length_run_number,101);
30 alpha_diff_many_runs = zeros(length_run_number,1);
31
32 plot_uni_ci_median_vector_old = zeros(length_run_number,101);
33 plot_uni_ci_upper_vector_old = zeros(length_run_number,101);
34 plot_uni_ci_lower_vector_old = zeros(length_run_number,101);
35 pfa_vector_true_curve_vector_old = zeros(length_run_number,101);
36 pd_vector_true_curve_vector_old = zeros(length_run_number,101);
37
38 aurc_median_many_runs = zeros(length_run_number,1);
```

```
 39 aurc_upper_many_runs = zeros(length_run_number,1);
 40 aurc_lower_many_runs = zeros(length_run_number,1);
 41 alpha_aurc_many_runs = zeros(length_run_number,1);
 42 true_aurc_many_runs = zeros(length_run_number,1);
 43
 44 run_number = 1;
 45
 46 run_number_start = 1;
 47 run_number_end = number_of_runs;
 48
 49 for run_number = run_number_start:run_number_end
 50
 51
 52     [plot_uni_ci_median,plot_uni_ci_upper,plot_uni_ci_lower,uni_conf_check_val,pfa_vector_true_curve,↙
pd_vector_true_curve,mse_uni,test_uni,alpha_diff,aurc_median,aurc_upper,aurc_lower,alpha_aurc,true_aurc] =↙
twobeta_run_nonempirical_2br...
 53         (run_number,number_of_target_samples,number_of_nontarget_samples,mean_target_1,std_target_1,mean_target_2,↙
std_target_2,...
 54     ratio_target,mean_nontarget_1,std_nontarget_1,mean_nontarget_2,std_nontarget_2,ratio_nontarget,↙
number_of_random_draws,ci_range);
 55
 56     aurc_median_many_runs(run_number) = aurc_median;
 57     aurc_upper_many_runs(run_number) = aurc_upper;
 58     aurc_lower_many_runs(run_number) = aurc_lower;
 59     alpha_aurc_many_runs(run_number) = alpha_aurc;
 60     true_aurc_many_runs(run_number) = true_aurc;
 61
 62     uni_conf_check_val_many_runs(run_number) = uni_conf_check_val;
 63
 64     mse_uni_many_runs(run_number) = mse_uni;
 65
 66     test_uni_many_runs(run_number) = test_uni;
 67     alpha_diff_many_runs(run_number) = alpha_diff;
 68
 69     plot_uni_ci_median_vector(run_number,:) = plot_uni_ci_median;
 70     plot_uni_ci_upper_vector(run_number,:) = plot_uni_ci_upper;
 71     plot_uni_ci_lower_vector(run_number,:) = plot_uni_ci_lower;
 72     pfa_vector_true_curve_vector(run_number,:) = pfa_vector_true_curve;
```

```
    pd_vector_true_curve_vector(run_number,:) = pd_vector_true_curve;

   % save two_beta_empiric_15_Sep_30_samp;

end;

run_number = 1

plot(0:.01:1,plot_uni_ci_median_vector(run_number,:),'k-.');
hold on;
plot(0:.01:1,plot_uni_ci_upper_vector(run_number,:),'k:');
hold on;
plot(0:.01:1,plot_uni_ci_lower_vector(run_number,:),'k:');
hold on;
plot(pfa_vector_true_curve_vector(run_number,:),pd_vector_true_curve_vector(run_number,:),'k');
legend('median','upper ci','lower ci','truth (underlying)');
title(['alpha est. = ',num2str(1-uni_conf_check_val_many_runs(run_number)),' run # = ',num2str(run_number),]);
axis equal;
axis([-.1 1.1 -.1 1.1]);
title(' ');
sort_mse_uni = sort(mse_uni_many_runs);

%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


[pfa_true_curve_mod,pd_true_curve_mod] = pfa_pd_to_hundredths_2br(pfa_vector_true_curve_vector(1,:),↙
pd_vector_true_curve_vector(1,:))

% alpha estimate per point

vector_check = [1:number_of_runs];

tot_runs_to_use = length(vector_check);

check_val_vector = zeros(101,1);
length_val_vector = zeros(101,1);
```

```
110
111 for index_a = 1:101
112     for index_b_wild = 1:tot_runs_to_use
113         index_b = vector_check(index_b_wild);
114         truth_compare_now = pd_true_curve_mod(index_a);
115         % metz process
116         low_compare_now = plot_uni_ci_lower_vector(index_b,index_a)
117         up_compare_now = plot_uni_ci_upper_vector(index_b,index_a)
118         length_now = up_compare_now-low_compare_now;
119         check_val = 1;
120         if low_compare_now <= truth_compare_now
121             if up_compare_now >= truth_compare_now
122                 check_val = 0;
123             end;
124         end;
125         check_val_vector(index_a) = check_val_vector(index_a) + check_val;
126         length_val_vector(index_a) = length_val_vector(index_a) + length_now;
127     end;
128     check_val_vector(index_a) = check_val_vector(index_a)./tot_runs_to_use;
129     length_val_vector(index_a) = length_val_vector(index_a)./tot_runs_to_use;
130 end;
131
132 test_pdf_vector = zeros(101,tot_runs_to_use);
133 cum_test_pdf_vector = zeros(size(test_pdf_vector));
134 Y_min = zeros(100,1);
135 Y_max = zeros(100,1);
136 min_error_vector = zeros(100,1);
137 max_error_vector = zeros(100,1);
138
139 for index = 2:100
140     test_pdf = binopdf(vector_check,length(vector_check),check_val_vector(index));
141     test_pdf_vector(index,:) = test_pdf;
142     for index_2 = 1:tot_runs_to_use
143         cum_test_pdf_vector(index,index_2) = sum(test_pdf_vector(index,1:index_2));
144     end;
145     [Y_min(index),I_min(index)] = min(abs(.05-cum_test_pdf_vector(index,:)));
146     [Y_max(index),I_max(index)] = min(abs(.95-cum_test_pdf_vector(index,:)));
147     min_error_vector(index) = I_min(index)./tot_runs_to_use;
```

```
148     max_error_vector(index) = I_max(index)./tot_runs_to_use;
149     index = index
150 end;
151
152 x_vals = 0:.01:1;
153 for index = 2:100
154     errorbar(x_vals(index),100.*(1-check_val_vector(index)),100.*(1-check_val_vector(index))-100.*(1-↙
min_error_vector(index)),100.*(1-max_error_vector(index))-100.*(1-check_val_vector(index)),'k');
155     hold on;
156 end;
157
158
159
160
161
162
163 %axis([-.05 1.05 68 102]);
164 line([-2 2],[90 90],'Color','k');
165 hold on;
166 plot(.01:.01:.99,(100.*(1-check_val_vector(2:100))),'k');
167 %hold on;
168 %axes('position',[-.05 1.05 68 102]);
169 % plot(pfa_true_curve_mod,pd_true_curve_mod,'k-.')
170 % % [AX,H1,H2] = plotyy(0:.01:1,(100.*(1-check_val_vector)),pfa_true_curve_mod,pd_true_curve_mod);
171 %
172 % set(H1,'LineStyle','--')
173 % set(H2,'LineStyle',':')
174 hold on;
175
176 mean_alpha = mean(1-uni_conf_check_val_many_runs);
177 %pbaspect([1 3 2]);
178 title(['mean auc alpha = ',num2str(mean_alpha)]);
179 axis([-.05 1.05 0 102]);
180
181
182
183
```

Appendix C-3
CEG/RSD Estimation and Confidence Interval Matlab Instructions
Beta Density Target and Non-target Model

A. Provide a set of target samples and non-target samples. Then estimate the CEG curve and RSD and associated confidence intervals based on these samples, (assuming a single beta density model).

1. Place the following files into a common directory.
   (c:\matlab_sv12\work\ceg\)

   beta_mean_c.m
   beta_mean_w_a_b_c.m
   conditioned_calc_2_c.m
   conf_error_new_w_return_c.m
   conf_error_new_weighted_c.m
   find_max_variance_c.m
   get_density_vals_c.m
   get_grid_points_c.m
   get_grid_points_closest_c.m
   get_grid_points_n_closest_c.m
   get_pd_pfa_matrix_10_c.m
   get_pd_pfa_pairs_pdfs2_c.m
   high_low_grid_weight_c.m
   mean_variance_to_pdf_2_c.m
   script_for_samples_c.m
   uni_ce_pdf_samples_c.m

2. Add the common directory to the Matlab path by using 'File / Set Path' option in Matlab menu, if the directory is not already in the path.

3. Execute the following in Matlab. An example is contained in 'script_for_samples_c.m'.

Enter (or load) a vector of target scores into the variable 'new_target_scores'.
Enter (or load) a vector of nontarget scores into the variable 'new_nontarget_scores'.

Execute the following matlab code to produce ROC curve and AUC value estimates with confidence intervals:

uni_ce_pdf_samples_c(new_target_scores,new_nontarget_scores,.95,.5);

Replace .95 by alternate confidence interval coverage if desired (e.g. .90, .80).

Replace .5 by prior probability of target.

To obtain the upper and lower confidence interval limit values for false alarm probabilities 0, .01, …, .99, 1, (rather than an on-screen plot), execute the following:
[ci_median, ci_upper, ci_lower, rsd_median, rsd_upper, rsd_lower] = …
uni_pdf_for_samples_c(new_target_scores,new_nontarget_scores,bound_value,prior_target);

ci_median – ROC curve estimate
ci_upper - Upper ROC curve confidence interval contour
ci_lower - Lower ROC curve confidence interval contour
rsd_median - AUC value estimate
rsd_upper - Upper AUC value confidence interval estimate
rsd_lower - Lower AUC value confidence interval estimate

B. Generate many sets of samples for selected underlying target and nontarget densities, and then obtain confidence intervals and estimates for the CEG curve / RSD value for each set of samples and compute confidence interval accuracy (e.g. alpha) among all sets. This process assumes a single beta model for target and non-target.

1. Place the following files into a common directory.

   For example: c:\matlab_sv12\work\roc\

   beta_mean_c.m
   beta_mean_w_a_b_c.m
   conditioned_calc_2_c.m
   conf_error_new_w_return_c.m
   conf_error_new_weighted_c.m
   find_max_variance_c.m
   get_density_vals_c.m
   get_grid_points_c.m
   get_grid_points_closest_c.m
   get_grid_points_n_closest_c.m
   get_pd_pfa_matrix_10_c.m
   get_pd_pfa_pairs_pdfs2_c.m
   high_low_grid_weight_c.m
   mean_variance_to_pdf_2_c.m
   script_CEG_RSD_CIs_with_coverage_accuracy.m
   run_choose_sample_ceg_c.m
   run_ceg_check_c.m
   sample_gen_choose_density_c.m
   sample_gen_uni_t_extend_c.m

2. Add the common directory to the Matlab path by using 'File / Set Path' option in Matlab menu.

3. Open the File 'script_CEG_RSD_CIs_with_coverage_accuracy.m'.

   Lines 8-21. Specify number of target samples, number of non-target samples, specify a beta density by mean and variance of assumed target beta density, mean and variance of assumed non-target beta density, number of runs (how many test runs are desired), and prior probability of target. Alternatively, specify the assumed target density and non-target density by the magnitude of the density at each of 1001 evaluation points (lines 24-27 provide an example).

   Evaluate Lines 1 through 83. [Note in Matlab this can be achieved by highlighting these line. Then right click to obtain a menu. Then choose 'Evaluate Selection'.] After each run, the full set of results can be saved per Line 81.

   ROC curve for a single run with confidence intervals:

   a. Form a plot of estimated ROC with confidence intervals [with true ROC] by Evaluating Lines 85-98. Note that run_number on line 92 may be adjusted to any run among the set specified in step 3 above.

   Obtain coverage for the full set of runs by Evaluating Lines 104-182. The mean alpha for RSD over many runs is displayed at the top of the plot.

```
% script_for_samples_c.m
new_target_scores = [0.2608; 0.3177; .3727; 0.3982; 0.4305; 0.4447; 0.4489; 0.4701; 0.4932; 0.5055; 0.5102;...
0.5129;...
    0.5213; 0.5366; 0.5524; 0.5743; 0.5906; 0.6163; 0.6232; 0.6367; 0.6579; 0.6590; 0.6669; 0.6731; 0.6773; 0.6814;...
0.7149;...
    0.7902; 0.8086; 0.8248];

new_nontarget_scores = [0.2189; 0.2569; 0.2719; 0.2995; 0.3040; 0.3336; 0.3500; 0.3622; 0.3710; 0.3793; 0.3932;...
0.4034; 0.4135;...
    0.4220; 0.4308; 0.4370; 0.4531; 0.4644; 0.5098; 0.5203; 0.5314; 0.5361; 0.5611; 0.5794; 0.5855; 0.6207; 0.6252;...
0.6309;...
    0.6540; 0.7053];



% load new_target_scores;
% load new_nontarget_scores;


% Specify the desired coverage:
bound_value = .95;
prior_target = .5;

uni_ce_pdf_samples_c(new_target_scores,new_nontarget_scores,bound_val,prior_target);




[uni_value_1, uni_value_2, uni_value_3, uni_value_4, uni_value_5, uni_value_6, uni_value_7] = uni_ce_pdf_samples_c...
(new_target_scores,new_nontarget_scores,bound_val,prior_target);

uni_ci_median = uni_value_1;
uni_ci_upper = uni_value_2;
uni_ci_lower = uni_value_3;
```

```
34 ce_median = uni_value_4;
35 ce_upper = uni_value_5;
36 ce_lower = uni_value_6;
37 dens_score = uni_value_7;
38
39 plot(.01:.01:.99,uni_ci_median(2:100),'k-.');
40 hold on;
41 plot(.01:.01:.99,uni_ci_upper(2:100),'k:');
42 hold on;
43 plot(.01:.01:.99,uni_ci_lower(2:100),'k:');
44 axis equal;
45 axis([-.1 1.1 -.1 1.1]);
46 xlabel('score');
47 ylabel('P(T|s)');
48 title(['rsd med. = ',num2str(ce_med_new),' rsd lower c.i. = ',num2str(ce_low_new),' rsd upper c.i. = ',num2str↙
(ce_up_new)]);
49
```

```
 1
 2 % script_CEG_RSD_CIs_with_coverage_accuracy.m
 3
 4 clear all;
 5 target_density = zeros(1000,1);
 6 nontarget_density = zeros(1000,1);
 7
 8 number_of_target_samples = 200;
 9 number_of_nontarget_samples = 200;
10
11 prior_target = .5;
12 % %
13 % % % Enter the beta density parameters
14 mean_target_gen_1 = .599;
15 std_target_gen_1 = .021.^.5;
16 mean_nontarget_gen_1 = .479;
17 std_nontarget_gen_1 = .023.^.5;
18
19 bound_val = .95;
20
21 number_of_runs = 1;
22
23 % optional:
24 % Enter target density and nontarget densities in place of above beta parameter values
25 % A beta pdf is used here as an example.
26 % target_density = betapdf(0:.001:1,3,1);
27 % nontarget_density = betapdf(0:.001:1,1,3);
28
29 length_run_number = 1;
30 plot_uni_ci_median_vector = zeros(length_run_number,101);
31 plot_uni_ci_upper_vector = zeros(length_run_number,101);
32 plot_uni_ci_lower_vector = zeros(length_run_number,101);
33 score_vector_true_curve_vector = zeros(length_run_number,1001);
34 pts_vector_true_curve_vector = zeros(length_run_number,1001);
35
36 rsd_median_many_runs = zeros(length_run_number,1);
37 rsd_upper_many_runs = zeros(length_run_number,1);
38 rsd_lower_many_runs = zeros(length_run_number,1);
```

```
 39 alpha_rsd_many_runs = zeros(length_run_number,1);
 40 true_rsd_many_runs = zeros(length_run_number,1);
 41 %
 42 ce_vector_list_matrix = zeros(length_run_number,12000);
 43
 44 run_number = 1;
 45
 46 run_number_start = 1;
 47 run_number_end = number_of_runs;
 48
 49 for run_number = run_number_start:run_number_end
 50
 51     if max(target_density) == 0
 52         variance_target_gen_1 = std_target_gen_1.^2;
 53         variance_nontarget_gen_1 = std_nontarget_gen_1.^2;
 54
 55         [plot_uni_ci_median,plot_uni_ci_upper,plot_uni_ci_lower,uni_conf_check_val,score_vector_true_curve,↙
pts_vector_true_curve,mse_uni,test_uni,alpha_diff,rsd_median,rsd_upper,rsd_lower,alpha_rsd,true_rsd] = run_ceg_check_c↙
(run_number,number_of_target_samples,number_of_nontarget_samples,...
 56             mean_target_gen_1,variance_target_gen_1,mean_nontarget_gen_1,variance_nontarget_gen_1,bound_val,↙
prior_target);
 57     end;
 58
 59     if max(target_density) > 0
 60         [plot_uni_ci_median,plot_uni_ci_upper,plot_uni_ci_lower,uni_conf_check_val,score_vector_true_curve,↙
pts_vector_true_curve,mse_uni,test_uni,alpha_diff,rsd_median,rsd_upper,rsd_lower,alpha_rsd,true_rsd] =↙
run_choose_sample_ceg_c(run_number,number_of_target_samples,number_of_nontarget_samples,...
 61             target_density,nontarget_density,bound_val,prior_target);
 62     end;
 63
 64     rsd_median_many_runs(run_number) = rsd_median;
 65     rsd_upper_many_runs(run_number) = rsd_upper;
 66     rsd_lower_many_runs(run_number) = rsd_lower;
 67     alpha_rsd_many_runs(run_number) = alpha_rsd;
 68     true_rsd_many_runs(run_number) = true_rsd;
 69     uni_conf_check_val_many_runs(run_number) = uni_conf_check_val;
 70
 71     mse_uni_many_runs(run_number) = mse_uni;
```

```
 72
 73     test_uni_many_runs(run_number) = test_uni;
 74
 75     plot_uni_ci_median_vector(run_number,:) = plot_uni_ci_median;
 76     plot_uni_ci_upper_vector(run_number,:) = plot_uni_ci_upper;
 77     plot_uni_ci_lower_vector(run_number,:) = plot_uni_ci_lower;
 78     score_vector_true_curve_vector(run_number,:) = score_vector_true_curve;
 79     pts_vector_true_curve_vector(run_number,:) = pts_vector_true_curve;
 80
 81     save ceg_curve_test_runs;
 82
 83 end;
 84
 85 run_number = 1
 86
 87 plot(.01:.01:.99,plot_uni_ci_median_vector(run_number,2:100),'k-.');
 88 hold on;
 89 plot(.01:.01:.99,plot_uni_ci_upper_vector(run_number,2:100),'k:');
 90 hold on;
 91 plot(.01:.01:.99,plot_uni_ci_lower_vector(run_number,2:100),'k:');
 92 hold on;
 93 % plot(pfa_vector_true_curve_vector(run_number,:),pd_vector_true_curve_vector(run_number,:),'k');
 94 plot(.001:.001:.999,pts_vector_true_curve_vector(run_number,2:1000),'k');
 95 legend('median','upper ci','lower ci','truth (underlying)');
 96 title(['alpha estimate = ',num2str(1-uni_conf_check_val_many_runs(run_number)),' run # = ',num2str(run_number)]);
 97 axis equal;
 98 axis([-.1 1.1 -.1 1.1]);
 99 title(' ');
100
101 alpha_check_vector = 1-uni_conf_check_val_many_runs;
102
103
104 [pfa_true_curve_mod,pd_true_curve_mod] = score_pts_to_hundredths_c(score_vector_true_curve_vector(1,:),↙
pts_vector_true_curve_vector(1,:))
105
106 % alpha estimate per point
107
108 tot_runs_to_use = number_of_runs;
```

```
109
110 vector_check = [1:1:length(tot_runs_to_use)]
111
112 check_val_vector = zeros(101,1);
113 length_val_vector = zeros(101,1);
114
115 for index_a = 1:101
116     for index_b_wild = 1:tot_runs_to_use
117         index_b = vector_check(index_b_wild);
118         truth_compare_now = pd_true_curve_mod(index_a);
119         % metz process
120         low_compare_now = plot_uni_ci_lower_vector(index_b,index_a)
121         up_compare_now = plot_uni_ci_upper_vector(index_b,index_a)
122         length_now = up_compare_now-low_compare_now;
123         check_val = 1;
124         if low_compare_now <= truth_compare_now
125             if up_compare_now >= truth_compare_now
126                 check_val = 0;
127             end;
128         end;
129         check_val_vector(index_a) = check_val_vector(index_a) + check_val;
130         length_val_vector(index_a) = length_val_vector(index_a) + length_now;
131     end;
132     check_val_vector(index_a) = check_val_vector(index_a)./tot_runs_to_use;
133     length_val_vector(index_a) = length_val_vector(index_a)./tot_runs_to_use;
134 end;
135
136 test_pdf_vector = zeros(101,tot_runs_to_use);
137 cum_test_pdf_vector = zeros(size(test_pdf_vector));
138 Y_min = zeros(100,1);
139 Y_max = zeros(100,1);
140 min_error_vector = zeros(100,1);
141 max_error_vector = zeros(100,1);
142
143 for index = 2:100
144     test_pdf = binopdf(vector_check,length(vector_check),check_val_vector(index));
145     test_pdf_vector(index,:) = test_pdf;
146     for index_2 = 1:tot_runs_to_use
```

```
147         cum_test_pdf_vector(index,index_2) = sum(test_pdf_vector(index,1:index_2));
148     end;
149     [Y_min(index),I_min(index)] = min(abs(.05-cum_test_pdf_vector(index,:)));
150     [Y_max(index),I_max(index)] = min(abs(.95-cum_test_pdf_vector(index,:)));
151     min_error_vector(index) = I_min(index)./tot_runs_to_use;
152     max_error_vector(index) = I_max(index)./tot_runs_to_use;
153     index = index
154 end;
155
156 x_vals = 0:.01:1;
157 for index = 2:100
158     errorbar(x_vals(index),100.*(1-check_val_vector(index)),100.*(1-check_val_vector(index))-100.*(1-↙
min_error_vector(index)),100.*(1-max_error_vector(index))-100.*(1-check_val_vector(index)),'k');
159     hold on;
160 end;
161
162
163
164
165
166
167 %axis([-.05 1.05 68 102]);
168 line([-2 2],[90 90],'Color','k');
169 hold on;
170 plot(.01:.01:.99,(100.*(1-check_val_vector(2:100))),'k');
171 %hold on;
172 %axes('position',[-.05 1.05 68 102]);
173 % plot(pfa_true_curve_mod,pd_true_curve_mod,'k-.')
174 % % [AX,H1,H2] = plotyy(0:.01:1,(100.*(1-check_val_vector)),pfa_true_curve_mod,pd_true_curve_mod);
175 %
176 % set(H1,'LineStyle','--')
177 % set(H2,'LineStyle',':')
178 hold on;
179 %pbaspect([1 3 2]);
180 mean_alpha = mean(1-uni_conf_check_val_many_runs);
181 title(['mean rsd alpha = ',num2str(mean_alpha)]);
182 axis([-.05 1.05 68 102]);
```

Appendix C-4
CEG/RSD Estimation and Confidence Interval Matlab Instructions
Two-Beta Mixture Target and Non-Target Density Model

A. Provide a set of target samples, non-target samples, and confidence bound value. Then compute the CEG curve and AUC value median estimates and confidence intervals based on these samples (assumes a two-beta mixture model).

1. Place the following files into a common directory.
   (For example: c:\matlab_sv12\work\)

   beta_mean_w_a_b_2bc.m
   combine_beta_pdf_2bc.m
   conditioned_calc_2_2bc.m
   conf_error_new_w_return_2bc.m
   conf_error_new_weighted_2bc.m
   find_max_variance_2bc.m
   mixture_pdf_2bc.m
   rand_two_beta_density_2bc.m
   sample_gen_bimodal_2bc.m
   score_pts_to_hundredths_2bc.m
   two_beta_script_for_given_samples_2bc.m
   twobeta_ceg_truth_not_known_2bc.m

2. Add the common directory to the Matlab path by using 'File / Set Path' option in Matlab menu, if the directory is not already in the path.

3. Execute the following in Matlab. An example is contained in 'two_beta_script_for_given_samples_2bc.m'

Enter (or load) a vector of target scores into the variable 'new_target_scores'.
Enter (or load) a vector of nontarget scores into the variable 'new_nontarget_scores'.

Execute the following matlab code to produce ROC curve and AUC value estimates with confidence intervals:

two_beta_ceg_truth_not_known(new_target_scores,new_nontarget_scores,10000,.95);

Replace 10000 by the desired number of random draws (lower numbers of draws decrease computational time). An approach is to begin with a low number of draws and gradually increase until convergence of confidence interval solution is observed.
Replace .95 by alternate confidence interval coverage if desired (e.g. .90, .80).

To obtain the upper and lower confidence interval limit values for scores 0, .01, …, .99, 1, (rather than an on-screen plot), execute the following:

[ci_median, ci_upper, ci_lower, rsd_median, rsd_upper, rsd_lower] = …
two_beta_ceg_truth_not_known(new_target_scores,new_nontarget_scores,10000,.95);

ci_median – ROC curve estimate
ci_upper - Upper ROC curve confidence interval contour
ci_lower - Lower ROC curve confidence interval contour
rsd_median - AUC value estimate
rsd_upper - Upper AUC value confidence interval estimate
rsd_lower - Lower AUC value confidence interval estimate

B. Generate many sets of samples for selected underlying target and nontarget densities, and then obtain confidence intervals and estimates for the CEG curve / RSD value for each set of samples and compute confidence interval accuracy (e.g. alpha) among all sets. This process assumes a single beta model for target and non-target.

1. Place the following files into a common directory.

   For example: c:\matlab_sv12\work\roc\

   beta_mean_w_a_b_2bc.m
   combine_beta_pdf_2bc.m
   conditioned_calc_2_2bc.m
   conf_error_new_w_return_2bc.m
   conf_error_new_weighted_2bc.m
   find_max_variance_2bc.m
   mixture_pdf_2bc.m
   rand_two_beta_density_2bc.m
   sample_gen_bimodal_2bc.m
   score_pts_to_hundredths_2bc.m
   two_beta_script_for_many_runs_2bc.m
   twobeta_run_nonempirical_2bc.m
   twobeta_unipdf_rsd_1000_2bc.m

2. Add the common directory to the Matlab path by using 'File / Set Path' option in Matlab menu.

3. Open the File 'two_beta_script_for_many_runs_2bc.m'.

   Lines 4-23. Specify number of target samples, number of non-target samples, specify a target beta density by the five parameters of an assumed beta density, specify non-target density by the five parameters of an assumed beta density, number of runs (how many test runs are desired), and prior probability of target. Also specify the number of random draws; this is a computational constraint, the number of draws selects how many grid points to evaluate for the target and non-target densities. An option is to begin at a number that executes quickly (e.g. 2000), then increase until observing convergence of computed confidence intervals.
   Evaluate Lines 1 through 71. [Note in Matlab this can be achieved by highlighting these line. Then right click to obtain a menu. Then choose 'Evaluate Selection'.] After each run, the full set of results can be saved per Line 69.

   ROC curve for a single run with confidence intervals:

   a. Form a plot of estimated ROC with confidence intervals [with true ROC] by Evaluating Lines 73-85. Note that run_number on line 92 may be adjusted to any run among the set specified in step 3 above.

   Obtain coverage for the full set of runs by Evaluating Lines 94-173. The mean alpha for RSD over many runs is displayed at the top of the plot.

```
 1
 2 new_target_scores = [0.7580; 0.7588; 0.7618; 0.7650; 0.7674; 0.7794; 0.7870; 0.7935;...
 3 0.7959; 0.8027; 0.8035; 0.8043; 0.8049; 0.8066; 0.8141; 0.8183; 0.8189; 0.8202;...
 4 0.8262; 0.8328; 0.8479; 0.8487; 0.8505; 0.8637; 0.8701; 0.8718; 0.8753; 0.8831;...
 5 0.8880; 0.8953];
 6
 7 new_nontarget_scores = [0.5988; 0.6331; 0.6396; 0.6642; 0.6652; 0.6688; 0.6734;...
 8 0.6736; 0.6764; 0.6802; 0.6954; 0.6990; 0.7024; 0.7059; 0.7109; 0.7126; 0.7132;...
 9 0.7162; 0.7185; 0.7196; 0.7231; 0.7286; 0.7347; 0.7362; 0.7644; 0.7647; 0.7785;...
10 0.7798; 0.7947; 0.8030];
11
12 ci_range = .90;
13 number_of_random_draws = 2000;
14
15 [uni_value_1, uni_value_2, uni_value_3, uni_value_4, uni_value_5, uni_value_6] = twobeta_roc_truth_not_known_2br↙
(new_target_scores,new_nontarget_scores,number_of_random_draws,ci_range);
16
17 uni_ci_median = uni_value_1;
18 uni_ci_upper = uni_value_2;
19 uni_ci_lower = uni_value_3;
20 aurc_median = uni_value_4;
21 aurc_upper = uni_value_5;
22 aurc_lower = uni_value_6;
23
24 plot(0:.01:1,uni_ci_median,'k-.');
25 hold on;
26 plot(0:.01:1,uni_ci_upper,'k:');
27 hold on;
28 plot(0:.01:1,uni_ci_lower,'k:');
29 hold on;
30 axis equal;
31 axis([-.1 1.1 -.1 1.1]);
32
33
34
```

```
 1
 2 clear all;
 3
 4 number_of_target_samples = 40;
 5 number_of_nontarget_samples = 40;
 6
 7 % First obtain the underlying target and clutter density:
 8 mean_target_1 = .599;
 9 std_target_1 = .021.^.5;
10 mean_target_2 = .6;
11 std_target_2 = .1;
12 ratio_target = .8;
13 %
14 mean_nontarget_1 = .479;
15 std_nontarget_1 = .023.^.5;
16 mean_nontarget_2 = .4;
17 std_nontarget_2 = .1;
18 ratio_nontarget = .85;
19
20 number_of_random_draws = 5000;
21 ci_range = .90;
22 prior_target = .5;
23 number_of_runs = 1;
24
25 length_run_number = 1;
26 plot_uni_ci_median_vector = zeros(length_run_number,101);
27 plot_uni_ci_upper_vector = zeros(length_run_number,101);
28 plot_uni_ci_lower_vector = zeros(length_run_number,101);
29 score_vector_true_curve_vector = zeros(length_run_number,101);
30 pts_vector_true_curve_vector = zeros(length_run_number,101);
31 alpha_diff_many_runs = zeros(length_run_number,1);
32
33 rsd_median_many_runs = zeros(length_run_number,1);
34 rsd_upper_many_runs = zeros(length_run_number,1);
35 rsd_lower_many_runs = zeros(length_run_number,1);
36 alpha_rsd_many_runs = zeros(length_run_number,1);
37 true_rsd_many_runs = zeros(length_run_number,1);
38
```

```
 39 run_number = 1;
 40
 41 run_number_start = 1;
 42 run_number_end = number_of_runs;
 43
 44 for run_number = run_number_start:run_number_end
 45
 46
 47     [plot_uni_ci_median,plot_uni_ci_upper,plot_uni_ci_lower,uni_conf_check_val,score_vector_true_curve,↙
pts_vector_true_curve,mse_uni,test_uni,rsd_median,rsd_upper,rsd_lower,alpha_rsd,true_rsd] =↙
twobeta_run_nonempirical_2bc...
 48         (run_number,number_of_target_samples,number_of_nontarget_samples,mean_target_1,std_target_1,mean_target_2,↙
std_target_2,...
 49     ratio_target,mean_nontarget_1,std_nontarget_1,mean_nontarget_2,std_nontarget_2,ratio_nontarget,↙
number_of_random_draws,ci_range,prior_target);
 50
 51     rsd_median_many_runs(run_number) = rsd_median;
 52     rsd_upper_many_runs(run_number) = rsd_upper;
 53     rsd_lower_many_runs(run_number) = rsd_lower;
 54     alpha_rsd_many_runs(run_number) = alpha_rsd;
 55     true_rsd_many_runs(run_number) = true_rsd;
 56
 57     uni_conf_check_val_many_runs(run_number) = uni_conf_check_val;
 58
 59     mse_uni_many_runs(run_number) = mse_uni;
 60
 61     test_uni_many_runs(run_number) = test_uni;
 62
 63     plot_uni_ci_median_vector(run_number,:) = plot_uni_ci_median;
 64     plot_uni_ci_upper_vector(run_number,:) = plot_uni_ci_upper;
 65     plot_uni_ci_lower_vector(run_number,:) = plot_uni_ci_lower;
 66     score_vector_true_curve_vector(run_number,:) = score_vector_true_curve;
 67     pts_vector_true_curve_vector(run_number,:) = pts_vector_true_curve;
 68
 69     save two_beta_many_runs_test_results;
 70
 71 end;
 72
```

```
run_number = 1

plot(.01:.01:.99,plot_uni_ci_median_vector(run_number,2:100),'k-.');
hold on;
plot(.01:.01:.99,plot_uni_ci_upper_vector(run_number,2:100),'k:');
hold on;
plot(.01:.01:.99,plot_uni_ci_lower_vector(run_number,2:100),'k:');
hold on;
plot(score_vector_true_curve_vector(run_number,2:100),pts_vector_true_curve_vector(run_number,2:100),'k');
legend('median','upper ci','lower ci','truth (underlying)');
title(['alpha est. = ',num2str(1-uni_conf_check_val_many_runs(run_number)),' run # = ',num2str(run_number),]);
axis equal;
axis([-.1 1.1 -.1 1.1]);
title(' ');
sort_mse_uni = sort(mse_uni_many_runs);

%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


[score_true_curve_mod,pts_true_curve_mod] = score_pts_to_hundredths_2bc(score_vector_true_curve_vector(1,:),pts_vector_true_curve_vector(1,:))

% alpha estimate per point

vector_check = [1:number_of_runs];

tot_runs_to_use = length(vector_check);

check_val_vector = zeros(101,1);
length_val_vector = zeros(101,1);

for index_a = 1:101
    for index_b_wild = 1:tot_runs_to_use
        index_b = vector_check(index_b_wild);
        truth_compare_now = pts_true_curve_mod(index_a);
        % metz process
```

```
        low_compare_now = plot_uni_ci_lower_vector(index_b,index_a)
        up_compare_now = plot_uni_ci_upper_vector(index_b,index_a)
        length_now = up_compare_now-low_compare_now;
        check_val = 1;
        if low_compare_now <= truth_compare_now
            if up_compare_now >= truth_compare_now
                check_val = 0;
            end;
        end;
        check_val_vector(index_a) = check_val_vector(index_a) + check_val;
        length_val_vector(index_a) = length_val_vector(index_a) + length_now;
    end;
    check_val_vector(index_a) = check_val_vector(index_a)./tot_runs_to_use;
    length_val_vector(index_a) = length_val_vector(index_a)./tot_runs_to_use;
end;

test_pdf_vector = zeros(101,tot_runs_to_use);
cum_test_pdf_vector = zeros(size(test_pdf_vector));
Y_min = zeros(100,1);
Y_max = zeros(100,1);
min_error_vector = zeros(100,1);
max_error_vector = zeros(100,1);

for index = 2:100
    test_pdf = binopdf(vector_check,length(vector_check),check_val_vector(index));
    test_pdf_vector(index,:) = test_pdf;
    for index_2 = 1:tot_runs_to_use
        cum_test_pdf_vector(index,index_2) = sum(test_pdf_vector(index,1:index_2));
    end;
    [Y_min(index),I_min(index)] = min(abs(.05-cum_test_pdf_vector(index,:)));
    [Y_max(index),I_max(index)] = min(abs(.95-cum_test_pdf_vector(index,:)));
    min_error_vector(index) = I_min(index)./tot_runs_to_use;
    max_error_vector(index) = I_max(index)./tot_runs_to_use;
    index = index
end;

x_vals = 0:.01:1;
for index = 2:100
```

```
148     errorbar(x_vals(index),100.*(1-check_val_vector(index)),100.*(1-check_val_vector(index))-100.*(1-↙
min_error_vector(index)),100.*(1-max_error_vector(index))-100.*(1-check_val_vector(index)),'k');
149     hold on;
150 end;
151
152
153
154
155
156
157 %axis([-.05 1.05 68 102]);
158 line([-2 2],[90 90],'Color','k');
159 hold on;
160 plot(.01:.01:.99,(100.*(1-check_val_vector(2:100))),'k');
161 %hold on;
162 %axes('position',[-.05 1.05 68 102]);
163 % plot(pfa_true_curve_mod,pd_true_curve_mod,'k-.')
164 % % [AX,H1,H2] = plotyy(0:.01:1,(100.*(1-check_val_vector)),pfa_true_curve_mod,pd_true_curve_mod);
165 %
166 % set(H1,'LineStyle','--')
167 % set(H2,'LineStyle',':')
168 hold on;
169 %pbaspect([1 3 2]);
170 mean_alpha = mean(1-uni_conf_check_val_many_runs);
171 %pbaspect([1 3 2]);
172 title(['mean rsd alpha = ',num2str(mean_alpha)]);
173 axis([-.05 1.05 0 102]);
174
175
176
177
```

Bibliography

[Agarwal *et al.*, 2005] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization Bounds for the Area under the ROC Curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

[Alsing, 2000] S.G. Alsing. *The Evaluation of Competing Classifiers. PhD dissertation*. Air Force Institute of Technology, 2000.

[Antelman, 1997] G. Antelman. *Elementary Bayesian Statistics*. Edward Elgar, 1997.

[Bamber, 1975] D. Bamber. The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Graph. *Journal of Mathematical Psychology*, 12:387–415, 1975.

[Barbieri and Berger, 2004] M.M. Barbieri and J.O. Berger. Optimal Predictive Model Selection. *The Annals of Statistics*, 32:870–897, 2004.

[Barkat, 1991] M. Barkat. *Signal Detection and Estimation.* Artech House, 1991.

[Bishop, 1995] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[Bolstad, 2004] W.M. Bolstad. *Introduction to Bayesian Statistics*. Wiley, 2004.

[Bos, 2002] C.S. Bos. A Comparison of Marginal Likelihood Computation Methods. *Tinbergen Institute Discussion Paper*, TI 2002-084/4, 2002.

[Bradley, 1997] A.P. Bradley. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[Broemeling, 2004] L.D. Broemeling. The Predictive Distribution and Area Under the ROC Curve. Technical report, The University of Texas M.D. Anderson Cancer Center Technical Report 013-04, 2004.

[Bryant, 2002] M.L. Bryant. *Manifold Recognition. PhD Dissertation*. Wright State University, 2002.

[Campbell and Ratnaparkhi, 1993] G. Campbell and M.V. Ratnaparkhi. An Application of Lomax Distributions in Receiver Operating Characteristic (ROC) Curve Analysis. *Communications in Statistics - Theory and Methods*, 22:1681–1697, 1993.

[Campbell, 1994] G. Campbell. Advances in Statistical Methodology for the Evaluation of Diagnostic and Laboratory Tests. *Statistics in Medicine*, 13:499–508, 1994.

[Carlin and Louis, 2000] B.P. Carlin and T.A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall/CRC, 2000.

[Carlin, 1992] J.B. Carlin. Meta-analysis for 2 x 2 Tables: A Bayesian Approach. *Statistics in Medicine*, 11:141–158, 1992.

[Carsten *et al.*, 2003] S. Carsten, S. Wesseling, T. Schink, and K. Jung. Comparison of Eight Computer Programs for Receiver-Operating Characteristic Analysis. *Clinical Chemistry*, 49:433–439, 2003.

[Casella and Berger, 2002] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury, Second edition, 2002.

[Centor, 1991] R.M. Centor. Signal Detectability: The Use of ROC Curves and their Analyses. *Medical Decision Making*, 11:102–106, 1991.

[Ceritoglu *et al.*, 2003] C. Ceritoglu, D. Bitouk, M.I. Miller, and H.A. Schmitt. Asymptotic Performance of ATR in Infrared Images. *Proceedings of the SPIE*, 5094:109–143, 2003.

[Claeskens *et al.*, 2003] G. Claeskens, B.-Y. Jing, L. Peng, and W. Zhou. Empirical Likelihood Confidence Regions for Comparison Distributions and ROC curves. *Canadian Journal of Statistics*, 31:173–190, 2003.

[Clyde and George, 2004] M. Clyde and E.I. George. Model Uncertainty. *Statistical Science*, 19:81–94, 2004.

[Clyde, 1999] M.A. Clyde. Bayesian Model Averaging and Model Search Strategies. In J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 6*, pages 157–185. Oxford, 1999.

[Cortes and Mohri, 2005] C. Cortes and M. Mohri. Confidence Intervals for the Area Under the ROC Curve. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005.

[Curiel, 2005] C. Curiel. Unmanned Systems RQ-4A Global Hawk High Altitude Endurance Unmanned Aerial Reconnaissance System - Northrup Grumman brochure. http://www.northrupgrumman.com/images/paris2005/pdfs/globalhawkbrochure.pdf, 2005.

[Dass and Jain, 2005] S.C. Dass and A.K. Jain. Effects of User Correlation on Sample Size Requirements. *Proceedings of the SPIE*, 5779:226–231, 2005.

[DeGroot and Schervish, 2002] M.H. DeGroot and M.J. Schervish. *Probability and Statistics*. Addison Wesley, third edition, 2002.

[DeLong *et al.*, 1988] E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44:837–845, 1988.

[DeVore, 2004] M.D. DeVore. Analytical Performance Evaluation of SAR ATR with Inaccurate or Estimated Models. *Proceedings of the SPIE*, 5427:407–417, 2004.

[Dodd and Pepe, 2003] L.E. Dodd and M.S. Pepe. Partial AUC Estimation and Regression. *Biometrics*, 59:614–623, 2003.

[Dorfman and Alf Jr., 1968] D.D. Dorfman and E. Alf Jr. Maximum Likelihood Estimation of Parameters of Signal Detection Theory - A Direct Solution. *Psychometrika*, 33:117–124, 1968.

[Dorfman and Alf Jr., 1969] D.D. Dorfman and E. Alf Jr. Maximum-Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals-Rating-Method Data. *Journal of Mathematical Psychology*, 6:487–496, 1969.

[Dorfman *et al.*, 1997] D.D. Dorfman, K.S. Berbaum, C.E. Metz, R.V. Lenth, J.A. Hanley, and H.A. Dagga. Proper Receiver Operating Characteristic Analysis: The Bigamma Model. *Academic Radiology*, 4(2):138–149, February 1997.

[Duda *et al.*, 2001] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, Second edition, 2001.

[Dukic and Gatsonis, 2003] V. Dukic and C. Gatsonis. Meta-analysis of Diagnostic Test Accuracy Assessment Studies with Varying Number of Thresholds. *Biometrics*, 49:936–946, 2003.

[Efron and Tibshirani, 1993] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[El-Jaroudi, 1990] A. El-Jaroudi. A New Error Criterion for Posterior Probability Estimation with Neural Nets. *International Joint Conference on Neural Networks*, pages 185–192, 1990.

[Eng, 2005] J. Eng. Receiver Operating Characteristic Analysis: A Primer. *Academic Radiology*, 12:909–916, 2005.

[Faraggi, 2003] D. Faraggi. Adjusting Receiver Operating Characteristic Curves and Related Indices for Covariates. *Statistician*, 51:179–192, 2003.

[Ferguson, 1967] T.S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, 1967.

[Garber *et al.*, 1994] A.M. Garber, R.A. Olshen, H. Zhang, and E.S. Venkatraman. Predicting High-Risk Cholestorol Levels. *International Statistical Review*, 62:203–228, 1994.

[Gelman *et al.*, 2004] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, second edition, 2004.

[Gokhale, 1975] D.V. Gokhale. Maximum Entropy Characterizations of Some Distributions. In G.P. Patil, S. Kotz, and J.K. Ord, editors, *Statistical Distributions in Scientific Work*, pages 299–304, 1975.

[Good, 1965] I.J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. The MIT Press, 1965.

[Green and Swets, 1988] D.M. Green and J.A. Swets. *Signal Detection Theory and Psychophysics*. Peninsula Publishing, Reprint; originally published in 1966 and 1974 edition, 1988.

[Greenhouse and Mantel, 1950] S.W. Greenhouse and N. Mantel. The Evaluation of Diagnostic Tests. *Biometrics*, 6:400–412, 1950.

[Gregory, 2005] P.C. Gregory. *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support*. Cambridge, 2005.

[Gustafson *et al.*, 2006] S.C. Gustafson, D.R. Parker, and R.K. Martin. Cardinal Interpolation. *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[Hahn and Shapiro, 1967] G.J. Hahn and S.S. Shapiro. *Statistical Models in Engineering*. Wiley, 1967.

[Hajian-Tilaki *et al.*, 1997] K.O. Hajian-Tilaki, J.A. Hanley, L. Joseph, and J.P. Collet. A Comparison of Parametric and Nonparametric Approaches to ROC Analysis of Quantitative Diagnostic Tests. *Medical Decision Making*, 17(1):94–102, 1997.

[Hall and Hyndman, 2003] P.G. Hall and R.J. Hyndman. Improved Methods for Bandwidth Selection when Estimating ROC Curves. *Statistics and Probability Letters*, 64:181–189, 2003.

[Hall *et al.*, 1991] P.S. Hall, T.K. Garland-Collins, R.S. Picton, and R.G. Lee. *Radar*. Brassey's, 1991.

[Hall *et al.*, 2004] P. Hall, R.J. Hyndman, and Y. Fan. Nonparametric Confidence Intervals for Receiver Operating Characteristic Curves. *Biometrika*, 91:743–750, 2004.

[Hammersley and Handscomb, 1964] J.M. Hammersley and D.C. Handscomb. *Monte Carlo Methods*. Methuen, 1964.

[Hanley, 1999] J.A. Hanley. Receiver Operating Characteristic (ROC) Curves. In Peter Armitage and Theodore Colton, editors, *Encyclopedia of Biostatistics*, pages 3738–3745. Wiley, 1999.

[Hellmich *et al.*, 1998] M. Hellmich, K.R. Abrams, D.R. Jones, and P.C. Lambert. A Bayesian Approach to a General Regression Model for ROC Curves. *Medical Decision Making*, 18:436–443, 1998.

[Hellmich *et al.*, 1999] M. Hellmich, K.R. Abrams, and A.J. Sutton. Bayesian Approaches to Meta-Analysis of ROC Curves. *Medical Decision Making*, 19:252–264, 1999.

[Hilgers, 1991] R.A. Hilgers. Distribution-Free Confidence-Bounds for ROC Curves. *Methods of Information in Medicine*, 30(2):96–101, April 1991.

[Hill *et al.*, 2003] J.M. Hill, M.E. Oxley, and K.W. Bauer. Receiver Operating Characteristic Curves and Fusion of Multiple Classifiers. *Proceedings of the Sixth International Conference of Information Fusion (Fusion 2003)*, 2:815–822, 2003.

[Hoeting *et al.*, 1999] J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volin. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417, 1999. Corrected version available at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

[Hogg and Craig, 1978] R.V. Hogg and A.T. Craig. *Introduction to Mathematical Statistics*. Prentice Hall, fourth edition, 1978.

[Hsieh and Turnbull, 1996] F. Hsieh and B.W. Turnbull. Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. *Annals of Statistics*, 24(1):25–40, 1996.

[Humphlett, 2004] J. Humphlett. Global Hawk Integrated Sensor Suite and Ground Segment - Raytheon brochure. Raytheon website: see: www.raytheon.com/products/Global Hawk Integrated Sensor Suite, 2004.

[Huntsberger, 1961] D.V. Huntsberger. *Elements of Statistical Inference*. Allyn and Bacon, 1961.

[Irvine *et al.*, 2002] J.M. Irvine, J.C. Mossing, D. Fitzgerald, K. Miller, and L.A. Westerkamp. Evaluation of Fusion-based ATR Technology. *Proceedings of the SPIE*, 4729:102–111, 2002.

[Jensen *et al.*, 2000] K. Jensen, H.-H. Muller, and H. Schafer. Regional Confidence Bands for ROC Curves. *Statistics in Medicine*, 19:493–509, 2000.

[Jordan *et al.*, 1999] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. *Learning in Graphical Models*, pages 105–161. MIT Press, 1999.

[Kagan *et al.*, 1973] A.M. Kagan, I. Linnik, and C. Rao. *Characterization Problems in Mathematical Statistics*. Wiley, 1973.

[Kass and Raftery, 1995] R.E. Kass and A.E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

[Korn and Korn, 2000] G.A. Korn and T.M. Korn. *Mathematical Handbook for Scientists and Engineers*. Dover, 2000.

[Larson *et al.*, 2002] R. Larson, R.P. Hostetler, B.H. Edwards, and D.E. Heyd. *Calculus with Analytic Geometry*. Houghton Mifflin, seventh edition, 2002.

[Lehmann and Casella, 1998] E.L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, second edition, 1998.

[Leonard and Hsu, 1999] T. Leonard and J.S.J. Hsu. *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. Cambridge, 1999.

[Li *et al.*, 2001] D. Li, M.R. Azimi-Sadjadi, and G.J. Dobeck. Comparison of Confidence Level of Different Classification Paradigms for Underwater Target Discrimination. *Proceedings of the SPIE*, 4394:1161–1172, 2001.

[Lindley, 1972] D.V. Lindley. *Bayesian Statistics, A Review*. SIAM, 1972.

[Linnet, 1987] K. Linnet. Comparison of Quantitative Diagnostic Tests: Type I Error, Power, and Sample Size. *Statistics in Medicine*, 6:147–158, 1987.

[Lloyd, 2002] C.J. Lloyd. Estimation of a Convex ROC Curve. *Statistics and Probability Letters*, 59:99–111, 2002.

[Lombard, 2003] F. Lombard. A New Approach to Determining the Precision and Bias of On-line Gauges. *Chemometrics and Intelligent Laboratory Systems*, 69:77–87, 2003.

[Lugosi and Pawlak, 1994] G. Lugosi and M. Pawlak. On the Posterior-probability Estimate of the Error Rate of Nonparametric Classification Rules. *IEEE Transactions on Information Theory*, 40:475–481, 1994.

[Lusted, 1971] L.B. Lusted. Signal Detectability and Medical Decision Making. *Science*, 171:1217–1219, 1971.

[Lusted, 1984] L.B. Lusted. ROC Recollected. *Medical Decision Making*, 4:131–135, 1984.

[Ma and Hall, 1993] G. Ma and W.J. Hall. Confidence Bands for Receiver Operating Characteristic Curves. *Medical Decision Making*, 13:191–197, 1993.

[MacKay, 1992a] D.J.C. MacKay. Bayesian Interpolation. *Neural Computation*, 4:415–447, 1992.

[MacKay, 1992b] D.J.C. MacKay. *Bayesian Methods for Adaptive Models. PhD Dissertation*. California Institute of Technology, 1992.

[MacKay, 2003] D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[Macskassy and Provost, 2004] S.A. Macskassy and F. Provost. Confidence Bands for ROC curves: Methods and an Empirical Study. In *Proceedings of the first workshop on ROC Analysis in AI (ROCAI-2004) at ECAI-2004, Spain*, 2004.

[Macskassy *et al.*, 2005] S.A. Macskassy, F. Provost, and S. Rosset. Pointwise ROC Confidence Bounds: An Empirical Evaluation. In *Proceedings of the ICML 2005 Workshop on ROC Analysis in Machine Learning, Bonn, Germany*, 2005.

[Madigan and Raftery, 1994] D. Madigan and A.E. Raftery. Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

[McNeil and Hanley, 1984] Barbara J. McNeil and James A. Hanley. Statistical Approaches to the Analysis of Receiver Operating Characteristic (ROC) Curves. *Medical Decision Making*, 4:137–150, 1984.

[Mendenhall *et al.*, 1990] W. Mendenhall, D.D. Wackerly, and R.L. Scheaffer. *Mathematical Statistics with Applications*. PWS-Kent, 1990.

[Messina and Meystel, 2004] E. Messina and A. Meystel, editors. *Performance Metrics for Intelligent Systems Workshop*. Sponsor: NIST, DARPA, IEEE, see http://www.isd.mel.nist.gov/PerMIS 2004, 2004.

[Metz and Pan, 1999] C.E. Metz and X.C. Pan. "Proper" Binormal ROC Curves: Theory and Maximum-likelihood Estimation. *Journal of Mathematical Psychology*, 43(1):1–33, March 1999.

[Metz *et al.*, 1998] C.E. Metz, B.A. Herman, and J.H. Shen. Maximum Likelihood Estimation of Receiver Operating Characteristic (ROC) Curves from Continuously-distributed Data. *Statistics in Medicine*, 17(9):1033–1053, 1998.

[Metz, 1984] C.E. Metz. Statistical Analysis of ROC Data in Evaluating Diagnostic Performance. In D. Herbert and R. Myers, editors, *Multiple Regression Analysis: Applications in the Health Sciences*, pages 365–384, 1984.

[Morgan, 1968] B.W. Morgan. *An Introduction to Bayesian Statistical Decision Processes*. Prentice-Hall, 1968.

[Mossing and Ross, 1998] J.C. Mossing and T.D. Ross. An Evaluation of SAR ATR Algorithm Performance Sensitivity to MSTAR Extended Operating Conditions. *Proceedings of the SPIE*, 3370:554–564, 1998.

[Mossman, 1995] D. Mossman. Resampling Techniques in the Analysis of Non-binormal ROC Data. *Medical Decision Making*, 15:358–366, 1995.

[Obuchowski and Lieber, 1998] N.A. Obuchowski and M.L. Lieber. Confidence Intervals for the Receiver Operating Characteristic Area in Studies with Small Samples. *Academic Radiology*, 5:561–571, 1998.

[O'Connor *et al.*, 2001] M.O. O'Connor, W. Remus, and K. Griggs. The Assymmetry of Judgemental Confidence Intervals in Time Series Forecasting. *International Journal of Forecasting*, 17:623–633, 2001.

[Olmstead, 1961] J.M.H. Olmstead. *Advanced Calculus*. Appleton-Century-Crofts, 1961.

[O'Malley *et al.*, 2001] A.J. O'Malley, K.H. Zou, J.R. Fielding, and C.M.C. Tempany. Bayesian Regression Methodology for Estimating a Receiver Operating Characteristic

Curve with Two Radiologic Applications: Prostate Biopsy and Spiral CT of Uretal Stones. *Academic Radiology*, 8:713–725, 2001.

[Papoulis, 1991] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.

[Parker *et al.*, 2005a] D.R. Parker, S.C. Gustafson, and T.D. Ross. Bayesian Confidence Intervals for ROC Curves. *IEE Electronics Letters*, 41:279–280, 2005.

[Parker *et al.*, 2005b] D.R. Parker, S.C. Gustafson, and T.D. Ross. Probability Densities and Confidence Intervals for Target Recognition Performance Metrics. *Proceedings of the SPIE*, 5808:373–382, May 2005.

[Parker *et al.*, 2005c] D.R. Parker, S.C. Gustafson, and T.D. Ross. Receiver Operating Characteristic and Confidence Error Metrics for Assessing the Performance of Automatic Target Recognition Systems. *Optical Engineering*, 44:097202, September 2005.

[Parker, 2005] D.R. Parker. Confidence Interval Generation Software for Receiver Operating Characteristic (ROC) Curves and Confidence Error Generation (CEG) Curves. Air Force Institute of Technology Technical Memorandum EN-TM-06-01, 2005.

[Patel *et al.*, 1976] J.K. Patel, C.H. Kapadia, and D.B. Owen. *Handbook of Statistical Distributions*. Marcel Dekker, 1976.

[Peng and Hall, 1996] F. Peng and W.J. Hall. Bayesian Analysis of ROC Curves Using Markov-chain Monte Carlo Methods. *Medical Decision Making*, 16:404–411, 1996.

[Platt *et al.*, 2000] R.W. Platt, J.A. Hanley, and H. Yang. Bootstrap Confidence Intervals for the Sensitivity of a Quantitative Diagnostic Test. *Statistics in Medicine*, 19:313–322, 2000.

[Poggio *et al.*, 2004] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General Conditions for Predictivity in Learning Theory. *Nature*, 428:419–422, 2004.

[Qiu and Le, 2001] P. Qiu and C. Le. ROC Curve Estimation Based on Local Smoothing. *Journal of Statistical Computation and Simulation*, 70:55–69, 2001.

[Raemer, 1997] H.R. Raemer. *Radar Systems Principles*. CRC Press, 1997.

[Raftery *et al.*, 2003] A.E. Raftery, F. Balabdaoui, T. Gneiting, and M. Polakowski. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Technical Report 440, University of Washington, December 2003.

[Robert, 2001] C.P. Robert. *The Bayesian Choice*. Springer, second edition, 2001.

[Ross and Minardi, 2004] T.D. Ross and M.E. Minardi. Discrimination and Confidence Error in Detector Reported Scores. *Proceedings of the SPIE*, 5427:342–353, 2004.

[Ross and Mossing, 1999] T.D. Ross and J.C. Mossing. The MSTAR Evaluation Methodology. *Proceedings of the SPIE*, 3721:705–713, 1999.

[Ross *et al.*, 1997] T.D. Ross, L.A. Westerkamp, E.G. Zelnio, and T.J. Burns. Extensibility and other Model-based ATR Evaluation Concepts. *Proceedings of the SPIE*, 3070:213–222, 1997.

[Ross *et al.*, 1998] T.D. Ross, S.W. Worrell, V.J. Velten, J.C. Mossing, and M.L. Bryant. Standard SAR ATR Evaluation Experiments using the MSTAR Public Release Data Set. *Proceedings of the SPIE*, 3370:566–573, 1998.

[Ross *et al.*, 1999] T.D. Ross, J.J. Bradley, L.J. Hudson, and M.P. O'Conner. SAR ATR - So What's The Problem? - An MSTAR Perspective. *Proceedings of the SPIE*, 3721:662–672, April 1999.

[Ross *et al.*, 2002] T.D. Ross, R.L. Dilsavor, J.C. Mossing, and L.A. Westerkamp. Performance Measures for Summarizing Confusion Matrices - The AFRL COMPASE approach. *Proceedings of the SPIE*, 4727:310–321, 2002.

[Ross, 2003] T.D. Ross. Accurate Confidence Intervals for Binomial Proportion and Poisson Rate Estimation. *Computers in Biology and Medicine*, 33:509–531, 2003.

[Rutter and Gatsonis, 2001] C.A. Rutter and C.A. Gatsonis. A Hierarchical Regression Approach to Meta-analysis of Diagnostic Test Accuracy Evaluations. *Statistics in Medicine*, 20:2865–2884, 2001.

[Schafer, 1994] H. Schafer. Efficient Confidence Bounds for ROC Curves. *Statistics in Medicine*, 13(15):1551–1561, 1994.

[Scharf, 1991] L.L. Scharf. *Statistical Signal Processing - Detection, Estimation, and Time Series Analysis*. Addison Wesley, July 1991.

[Schervish, 1995] M.J. Schervish. *Theory of Statistics*. Springer-Verlag, 1995.

[Schmitt, 1969] S.A. Schmitt. *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics*. Addison-Wesley, 1969.

[Schubert *et al.*, 2005] C.M. Schubert, M.E. Oxley, and K.W. Bauer Jr. Quantifying the Performance of Fused Correlated Multiple Classifiers. *Proceedings of the SPIE*, 5809:390–401, 2005.

[Shanmugan and Breipohl, 1988] K.S. Shanmugan and A.M. Breipohl. *Random Signals, Detection, Estimation and Data Analysis*. Wiley, 1988.

[Silverman, 1986] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[Simpson *et al.*, 1989] E.V. Simpson, R.E. Ideker, K.L. Lee, and W.M. Smith. Computing ROC Curve Confidence Intervals for Cardiac Activation Detectors. *IEEE Engineering in Medicine and Biology Society 11th Annual International Conference*, 1989.

[Smith *et al.*, 1995] T.C. Smith, D.J. Spiegelhalter, and A. Thomas. Bayesian Approaches to Random-Effects Meta-analysis: A Comparative Study. *Statistics in Medicine*, 14:2685–2699, 1995.

[Smith *et al.*, 1996] P.J. Smith, T.J. Thompson, M.M. Engelgau, and W.H. Herman. A Generalized Linear Model for Analysing Receiver Operating Characteristic Curves. *Statistics in Medicine*, 15:323–333, 1996.

[Sorribas *et al.*, 2002] A. Sorribas, J. March, and J. Trujillano. A New Parametric Method Based on S-distributions for Computing Receiver Operating Characteristic Curves for Continuous Diagnostic Tests. *Statistics in Medicine*, 21(9):1213–1235, May 2002.

[Stark and Woods, 1986] H. Stark and J.W. Woods. *Probability, Random Processes and Estimation Methods for Engineers*. Prentice-Hall, 1986.

[Swets, 1988] J.A. Swets. Measuring the Accuracy of Diagnostic Systems. *Science*, 240:1285–1293, 1988.

[Swetz and Pickett, 1982] J.A. Swetz and R.M. Pickett. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press, 1982.

[Thing, 2002] L. Thing. *Encyclopedia of Technology Terms*. Que, 2002.

[Thorsen and Oxley, 2004] S.N. Thorsen and M.E. Oxley. A Category Theory Description of Multisensor Fusion. *Proceedings of the SPIE*, 5434:261–269, 2004.

[Tilbury *et al.*, 2000] J.B. Tilbury, P.W. VanEetvelt, J.M. Garibaldi, J.S. Curnow, and E.C. Ifeachor. Receiver Operating Characteristic Analysis for Intelligent Medical Systems - A New Approach for Finding Confidence Intervals. *IEEE Transactions on Biomedical Engineering*, 47(7):952–963, 2000.

[Tilbury *et al.*, 2003a] J.B. Tilbury, P.W.J. Van Eetvelt, J.S.H. Curnow, and E.C. Ifeachor. Objective Evaluation of Intelligent Medical Systems using a Bayesian Approach to Analysis of ROC Curves. In *Fifth International Conference on Neural Networks and Expert Systems in Medicine and Healthcare*, 2003.

[Tilbury *et al.*, 2003b] J.B. Tilbury, P.W.J. Van Eetvelt, J.S.H. Curnow, and E.C. Ifeachor. Objective Evaluation of Intelligent Medical Systems using a Bayesian Approach to Analysis of ROC Curves. *Unpublished*, 2003. see http:/www.tech.plym.ac.uk/spmc/staff/jtilbury/.

[Tilbury, 2002] J.B. Tilbury. *Evaluation of Intelligent Medical Systems*. PhD thesis, University of Plymouth, England, 2002.

[Tomasi, 2004] C. Tomasi. Past Performance and Future Results. *Nature*, 428:378–378, 2004.

[Tosteson and Begg, 1988] A.N. Tosteson and C.B. Begg. A General Regression Methodology for ROC Curve Estimation. *Medical Decision Making*, 8:204–215, 1988.

[Trochim, 2005] W.M. Trochim. *The Research Methods Knowledge Base*. http://trochim.human.cornell.edu/kb/statistics.htm, Second edition, 2005.

[VanTrees, 1968] H.L. VanTrees. *Detection, Estimation, and Modulation Theory, Part I Detection, Estimation, and Linear Modulation Theory*. Wiley, 1968.

[Wickens, 2002] T.D. Wickens. *Elementary Signal Detection Theory*. Oxford University Press, 2002.

[Wieand *et al.*, 1989] S. Wieand, M.H. Gail, B.R. James, and K.L. James. A Family of Nonparametric Statistics for Comparing Diagnostic Markers with Paired or Unpaired Data. *Biometrika*, 76:585–592, 1989.

[Wise *et al.*, 2004] A.R. Wise, D. Fitzgerald, and T.D. Ross. The Adaptive SAR ATR Problem Set (AdaptSAPS). *Proceedings of the SPIE*, 5427:366–375, 2004.

[Woodworth, 2004] G.G. Woodworth. *Biostatistics: A Bayesian Introduction*. Wiley, 2004.

[Yaniv and Foster, 1997] I. Yaniv and D.P. Foster. Precision and Accuracy of Judgmental Estimation. *Journal of Behavioral Decision Making*, 10:21–32, 1997.

[Yousef *et al.*, 2005] W.A. Yousef, R.F. Wagner, and M.H. Loew. Estimating the Uncertainty in the Estimated Mean Area Under the ROC Curve of a Classifier. *Pattern Recognition Letters*, 26:2600–2610, 2005.

[Zelnio *et al.*, 2005] E.G. Zelnio, T.D. Ross, and M.L. Bryant. A Demonstration of the Confuser and Likelihood Modeling Benefits for Target Detection in SAR Imagery. *Proceedings of the SPIE*, 5808:395–406, 2005.

[Zhou and Qin, 2005] X.-H. Zhou and G. Qin. Improved Confidence Intervals for the Sensitivity at a Fixed Level of Specificity of a Continuous-scale Diagnostic Test. *Statistics in Medicine*, 24:465–477, 2005.

[Zhou, 1996] X.H. Zhou. Empirical Bayes Combination of Estimated Areas under ROC Curves Using Estimating Equations. *Medical Decision Making*, 16:24–28, 1996.

[Zou and O'Malley, 2005] K.H. Zou and A.J. O'Malley. A Bayesian Hierarchical Non-Linear Regression Model in Receiver Operating Characteristic Analysis of Clustered Continuous Diagnostic Data. *Biometrical Journal*, 47:417–427, 2005.

[Zou *et al.*, 1997] K.H. Zou, W.J. Hall, and D.E. Shapiro. Smooth Non-parametric Receiver Operating Characteristic (ROC) Curves for Continuous Diagnostic Tests. *Statistics in Medicine*, 16:2143–2156, 1997.

[Zou *et al.*, 2004] K.H. Zou, W.M. Wells III, R. Kikinis, and S.K. Warfield. Three Validation Metrics for Automated Probabilistic Image Segmentation of Brain Tumors. *Statistics in Medicine*, 23:1259–1282, 2004.

[Zweig and Campbell, 1993] M.H. Zweig and G. Campbell. Receiver Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry*, 39:561–577, 1993.

Vita

David R. Parker, Maj, USAF, is from Florence, New Jersey. In June 1990, Maj Parker accepted an appointment to the United States Air Force Academy (USAFA), Colorado Springs, Colorado. Maj Parker graduated from USAFA in 1994 with a B.S. in Electrical Engineering. Following graduation, Maj Parker completed Basic Communications Officer Training School at Keesler AFB, MS. From late 1994 to August 1998, Maj Parker was assigned as a communications officer at Fort George G. Meade, MD. In May 1998, Maj Parker was awarded a M.E.S. degree by Loyola College, Baltimore Maryland. From August 1998 to August 2002, Maj Parker was a project manager at the Space and Missile Systems Center, Los Angeles AFB, CA, and led a government team in defining and managing multiple advanced communications projects to support acquisitions for critical Department of Defense satellite communications systems. Maj Parker was selected for the AFIT Ph.D. program in 2001 and began studies at AFIT in September 2002. Maj Parker's Specialty Sequence at AFIT is Target Recognition/Signal Processing, and his Minor is in Electro-Optics. At AFIT, Maj Parker is also an Air Force Intermediate Developmental Education (IDE) student, and he completed Air Command and Staff College by Seminar in March 2005. Maj Parker's next assignment is at the Air Force Research Laboratory, Sensor's Directorate, Wright-Patterson AFB, OH. Maj Parker is a member of SPIE, IEEE, Eta Kappa Nu, and Tau Beta Pi.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
23-03-2006	Doctoral Dissertation	Jul 2003 – Mar 2006

4. TITLE AND SUBTITLE

UNCERTAINTY ESTIMATION FOR TARGET DETECTION SYSTEM DISCRIMINATION AND CONFIDENCE PERFORMANCE METRICS

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Parker, David R., Major, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management
2950 Hobson Way, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/DS/ENG/06-01

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Timothy D. Ross
Air Force Research Laboratory
AFRL/SNA
WPAFB OH 45433
937-255-5668, e-mail: timothy.ross@wpafb.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This research uses a Bayesian framework to develop probability densities for target detection system performance metrics. The metrics include the receiver operating characteristic (ROC) curve and the confidence error generation (CEG) curve. The ROC curve is a discrimination metric that quantifies how well a detection system separates targets and non-targets, and the CEG curve indicates how well the detection system estimates its own confidence. The degree of uncertainty in these metrics is a concern that previous research has not adequately addressed. This research formulates probability densities of the metrics and characterizes their uncertainty using confidence bands. Additional statistics are obtained that verify the accuracy of the confidence bands. Methods for the generation and characterization of the probability densities of the metrics are specified and demonstrated, where the initial analysis employs beta densities to model target and non-target samples of detection system output. For given target and non-target data, given functional forms of the data densities (such as beta density forms), and given prior densities of the form parameters, the methods developed here provide exact performance metric probability densities. Computational results compare favorably with existing approaches in cases where they can be applied; in other cases the methods developed here produce results that existing approaches can not address.

15. SUBJECT TERMS

Automatic target recognition, Performance metrics, Receiver operating characteristic, ROC curves, Target detection

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Steven C. Gustafson (AFIT/ENG)
U	U	U	UU	252	19b. TELEPHONE NUMBER *(Include area code)* 937-255-3636, e-mail: steven.gustafson@afit.edu